PLOS ONE

RESEARCH ARTICLE

Comparative genome analysis indicates high evolutionary potential of pathogenicity genes in *Colletotrichum tanaceti*

Ruvini V. Lelwala[1], Pasi K. Korhonen[1], Neil D. Young[1], Jason B. Scott[2], Peter K. Ades[3], Robin B. Gasser[1], Paul W. J. Taylor[1]*

1 Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, Victoria, Australia, **2** Tasmanian Institute of Agriculture, University of Tasmania, Burnie, Tasmania, Australia, **3** Faculty of Science, The University of Melbourne, Parkville, Victoria, Australia

* paulwjt@unimelb.edu.au

OPEN ACCESS

Citation: Lelwala RV, Korhonen PK, Young ND, Scott JB, Ades PK, Gasser RB, et al. (2019) Comparative genome analysis indicates high evolutionary potential of pathogenicity genes in *Colletotrichum tanaceti*. PLoS ONE 14(5): e0212248. https://doi.org/10.1371/journal.pone.0212248

Editor: Craig Eliot Coleman, Brigham Young University, UNITED STATES

Received: January 24, 2019

Accepted: May 2, 2019

Published: May 31, 2019

Data Availability Statement: Data are available from NCBI (accession number PJEX00000000).

Funding: This research was funded by Botanical Resources Australia - Agricultural Services, Pty. Ltd- (https://www.botanicalresources.com/). Ruvini V. Lelwala received Melbourne International Fee Remission Scholarship and Melbourne International Research Scholarship from the University of Melbourne, Australia (https://www.unimelb.edu.au/). ND Young was supported by

Abstract

Colletotrichum tanaceti is an emerging foliar fungal pathogen of commercially grown pyrethrum (*Tanacetum cinerariifolium*). Despite being reported consistently from field surveys in Australia, the molecular basis of pathogenicity of *C. tanaceti* on pyrethrum is unknown. Herein, the genome of *C. tanaceti* (isolate BRIP57314) was assembled *de novo* and annotated using transcriptomic evidence. The inferred putative pathogenicity gene suite of *C. tanaceti* comprised a large array of genes encoding secreted effectors, proteases, CAZymes and secondary metabolites. Comparative analysis of its putative pathogenicity gene profiles with those of closely related species suggested that *C. tanaceti* likely has additional hosts to pyrethrum. The genome of *C. tanaceti* had a high repeat content and repetitive elements were located significantly closer to genes inferred to influence pathogenicity than other genes. These repeats are likely to have accelerated mutational and transposition rates in the genome, resulting in a rapid evolution of certain CAZyme families in this species. The *C. tanaceti* genome showed strong signals of Repeat Induced Point (RIP) mutation which likely caused its bipartite nature consisting of distinct gene-sparse, repeat and A-T rich regions. Pathogenicity genes within these RIP affected regions were likely to have a higher evolutionary rate than the rest of the genome. This "two-speed" genome phenomenon in certain *Colletotrichum* spp. was hypothesized to have caused the clustering of species based on the pathogenicity genes, to deviate from taxonomic relationships. The large repertoire of pathogenicity factors that potentially evolve rapidly due to the plasticity of the genome, indicated that *C. tanaceti* has a high evolutionary potential. Therefore, *C. tanaceti* poses a high-risk to the pyrethrum industry. Knowledge of the evolution and diversity of the putative pathogenicity genes will facilitate future research in disease management of *C. tanaceti* and other *Colletotrichum* spp.

NHMRC Career Development Fellowship - APP1109829 (https://nhmrc.gov.au/). PK Korhonen was supported by NHMRC Early Career Research Fellowship - APP1127033 (https://nhmrc.gov.au/). The funders had no role in study design, data collection and analysis, decision to publish or Preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Plant pathogens cause diseases world-wide that have devastating economic, social and ecological consequences [1]. Fungi are among the dominant causal agents of plant diseases [2] and the genus *Colletotrichum* has been ranked among the top-ten most important fungal plant pathogens [3]. Many *Colletotrichum* species are known to cause major economic losses globally, and have been extensively used in the study of the molecular and cellular bases of fungal pathogenicity [4]. The publication of 25 whole genome sequences of *Colletotrichum* species has significantly improved understanding of the biology, genetics and evolution of this genus [5–11]. However, a large research gap still exists with this ever-expanding genus consisting of more than 200 accepted species [12] and 14 major species complexes [13, 14]. The availability of only one genome of a member of the destructivum complex, *C. higginsianum*, [5, 15] has constrained comparative studies within and among species complexes. Insights into the genomic organization and the pathogenicity gene repertoire of other *Colletotrichum* species in the destructivum complex therefore, will significantly expand the knowledge base of this important genus.

Colletotrichum tanaceti, a member of the destructivum complex [16], is an emerging foliar fungal pathogen [17] of Dalmatian pyrethrum (*Tanacetum cinerariifolium*). Pyrethrum is commercially cultivated as a source of the natural insecticide pyrethrin [18]. *Colletotrichum tanaceti* has been consistently reported in Australian field surveys of the crop [19] since 2012 [17] and causes leaf anthracnose, with black, water-soaked, sunken lesions [17]. Due to its hemibiotrophic lifestyle, characteristic symptoms of *C. tanaceti* are not evident on leaves until around 120 hours after infection [17, 20], when it switches from biotrophy to necrotrophy. A significant reduction in green leaf area occurs usually 10 days after infection [17]. This suggests a rapid disease cycle for *C. tanaceti* in pyrethrum and, given its aggressiveness, the potential for serious crop damage. The molecular basis of pathogenicity of *C. tanaceti*, which includes the pathogenicity genes and their evolution, has not been studied. The genome sequence of an emerging plant pathogen such as *C. tanaceti* is a good source for identifying putative genes associated with the pathogen life cycle, pathogenicity and virulence. Effectors [21], proteases [22], and carbohydrate active enzymes (CAZymes) [23] are such important gene categories in fungal pathogenesis. Furthermore, secondary metabolites and transporters, *P450s* and transcription factors [24] associated with biosynthesis of secondary metabolites are also important pathogenicity factors. Fungal mitogen activated protein (MAP) kinase pathways regulate the cascade of reactions that respond to various environmental stresses and are also important factors determining pathogenicity and virulence [25]. Draft genomes of many fungal pathogens have been used to infer genes involved in pathogenicity with a high accuracy [26, 27] using homology searches against curated databases [28, 29] and *de novo* inference using bioinformatics tools [21, 30]. Identification of putative pathogenicity genes of *C. tanaceti* is fundamental for assessing the present risk of the pathogen, for future studies of functional validation and ultimately for economic disease management.

The genome of a pathogen is also a good source for assessing evolutionary potential [31–33] as the adaptive evolution increases with the plasticity of the genome [34, 35]. In filamentous plant pathogens, repeat-rich gene-sparse genomic regions tend to harbor genes that are involved in pathogenicity and host adaptation [35] and evolve at higher rates than the rest of the genome giving rise to "two-speed genomes" [36]. Repeat-induced-point mutation (RIP) is a fungal-specific mechanism for limiting transposon proliferation below destructive levels [37]. Over time, RIP can cause the formation of A-T rich regions and is a mechanism facilitating two-speed fungal genomes [38–41]. The genome of *C. tanaceti* can be used to identify such

genomic architecture and their relationship to pathogenicity genes, in order to assess the plasticity and thereby the evolutionary potential.

Comparative genomics has enabled inference of patterns of speciation, pathogenesis and host determination within *Colletotrichum* lineages [42]. These studies have indicated that the gain and loss of putative pathogenicity gene families in *Colletotrichum* genomes are important determinants of host specificity and pathogenic adaptation of these species [7, 9–11, 43]. *Colletotrichum tanaceti* has only been reported from pyrethrum in Australia but may have crossed over from another host plant species. However, cross-host pathogenicity has not yet been assessed and the potential host range of the pathogen is currently unknown. Comparison of putative pathogenicity gene repertoires of *Colletotrichum* species from different species complexes and species closely related to the genus *Colletotrichum* may provide insights into evolution of pathogenicity gene and the host range of *C. tanaceti*. Therefore, combined genomics and comparative genomics analyses can provide sound means of assessing the current and future risks posed by *C. tanaceti*.

In order to achieve the major goal of evaluating the potential risk to the pyrethrum industry form *C. tanaceti*, the aims of this study were to: 1) infer the pathogenicity gene suite of *C. tanaceti*; 2) infer the host range of *C. tanaceti*; and 3) assess the evolutionary potential of pathogenicity genes of *C. tanaceti*.

Materials and methods

Sequencing and *de novo*-assembly of the genome of *C. tanaceti*

Fungal strain. The ex-holotype of *C. tanaceti* strain BRIP57314 (CBS 132693 = UM01) [17] was acquired from the culture collection of BRIP (Plant Pathology Herbarium, Department of Primary Industries, Queensland, Australia). This isolate was propagated on potato dextrose agar (PDA; Sigma Aldrich, St. Louis, USA) and incubated at 24˚C using a 12 h:12 h light:dark photoperiod. Genomic DNA was isolated using a modified CTAB protocol [44]. The integrity and quantity of DNA was confirmed by 1.5% agarose gel electrophoresis and a nanodrop spectrophotometer (Thermo Fisher Scientific, Waltham, USA).

Genome sequencing and assembly. Genomic DNA was fragmented using a Covaris ultrasonicator (Covaris Inc., Massachusetts, USA) to achieve an average fragment length of 532 base pairs (bp). A genomic DNA library with an average insert size of 420 bp was constructed using the KAPA Hyper Prep Library Preparation Kit [45] and was paired-end sequenced (2×300 bp reads) using the Illumina Miseq platform (San Diego, USA). The raw reads were filtered for low quality nucleotides and adapters using Trimmomatic [46] (Phred score-33, leading-3, trailing-6, slidingwindow-4:15, minlen-36) to retain 22,871,341 sequences and were profiled using KAT [47]. Filtered reads were then assembled using DISCOVAR *de novo* [48]. The completeness of the assembly was assessed with the Sordaromyceta_*odb9* gene set [49] using the program Benchmarking Universal Single-Copy Orthologs (BUSCO v2) [49] in the Genomics Virtual Laboratory platform [50].

Prediction of repetitive elements. Species-specific repeats were first inferred using the program RepeatModeler [51], in which the programs RECON [52] and RepeatScout [53] were used. Long terminal repeats (LTRs) were predicted using the program LTR_Finder [54]. The program RepeatMasker v4.0.5 [55] was employed to mask resulting species-specific repeats and LTRs; and applied the program Tandem Repeat Finder (TRF) [56] and the database Repbase v.17.02 [57] to predict and mask interspersed and simple repeats. All repeats predicted were combined using ProcessRepeats command in RepeatMasker.

RNA sequencing

Inoculation of pyrethrum leaves. Pyrethrum leaves were inoculated using the leaf-sandwich method [58, 59] by placing a fungal 'mat' between two pyrethrum leaves in a petridish. Each petri dish was sealed with parafilm and incubated at 24˚C with a 12 h-photoperiod. Induced mycelia were harvested at 6, 24 and 48 h after inoculation, and total RNA was extracted using the RNeasy Plant Mini kit (Qiagen, Australia) following the manufacturer's instructions. Total RNA was extracted from the saprobic stage (1-week-old cultures growing on potato dextrose agar). Contaminating genomic DNA was removed from RNA samples by Ambion DNase I (Thermo Fisher Scientific, USA) treatment; the integrity and quantity of total RNA was confirmed by 1% agarose gel electrophoresis and the Experion automated electrophoresis system (Biorad Laboratories, Australia).

RNA libaries were prepared using both E7530L and E&335L NEBNext Ultra RNA Library Prep Kits (New England Biolabs, USA) to generate fragment sizes of 351–371 bp. The transcriptome was paired-end sequenced (2 × 150 bp reads) on the Illumina Hiseq 2500 platform (San Diego, USA). Raw reads were trimmed for quality using Trimmomatic [46] (leading-25, trailing-25, slidingwindow-4:25, minlen-40) to retain between 17,935,938–18,761,773 sequences for each library and profiled using FastQC [60].

Gene prediction

Genes were first predicted using the MAKER3 v3.0.0-beta [61], in which both the transcriptomic data from *C. tanaceti* and the proteomic and *ab initio* gene predictions from *C. graminicola;*[43] and *C. higginsianum;* [43, 62] were combined into a consensus prediction. In brief, transcriptomic RNAseq reads of *C. tanaceti* were assembled into transcripts in both *de novo* and genome-guided modes of the program Trinity v2.2.0 [63]. In genome guided assembly, reads were mapped onto the genome using the program TopHat2 v2.1.0 [64]. Genome guided and *de novo* transcriptomic assemblies were combined, redundancy (99% similarity) was removed using the program cd-hit-est [65, 66] and resulting transcripts were filtered for full-length open reading frames (ORFs) using the program Transdecoder [63]. Resulting full-length transcripts were further reduced to 80% similarity using the program cd-hit-est and checked for splicing sites. These high quality transcripts were then used as a training set for *ab initio* gene prediction programs AUGUSTUS v3.1 [67] and SNAP v6.7 [68] and GENEMARK v4.2.9 [69]. Evidence data from assembled transcriptomes (with 99% redundancy using cd-hit-est) and the proteomes were provided to Maker3. The predicted genes (length of conceptually translated protein $\geq$ 30 amino acids) were further clustered using the *k*-means clustering algorithm [70] with following metrics: 1) Maker3 annotation edit distance (AED); 2) number of exons in the mRNA; 3) length of translated protein sequence; 4) fraction of exons that overlap transcript alignment; 5) fraction of exons that overlap transcript and protein alignment; 6) fraction of splice sites confirmed by a SNAP prediction from Maker3; 7) percentage for repeat overlap with gene-, exon- and CDS-sequence; 8) size of the inferred orthologous group the gene belongs to using OrthoMclv2.0.9 [71]; and 9) presence of functional annotation (see Functional annotation of the *C. tanaceti* genome section below). Resulting clusters with transposons and *ab initio* gene predictions with no transcriptome or proteome support were removed.

Functional annotation of the *C. tanaceti* genome

Putative coding regions were subjected to protein homology searches against the NCBI (nr) and Swiss-Prot database using BLAST v 2.7.1 (E-value of $\leq$ 1e-8) [72]. Conserved protein domains and gene ontology (GO) terms were assigned to predicted proteins using

InterProScan 5 [73]. Additionally, Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology (KO) terms were assigned to predicted proteins using the Blastkoala search engine [74]. Assigned KO terms were used to generate *C. tanaceti* pathway maps using KEGG mapper [75]. Putative genes of *C. tanaceti* with functional annotations were subjected to species-specific gene enrichment analysis on the DAVID functional annotation database tool [76, 77] and using *C. graminicola* as the reference species.

Comparison to related taxa

The genome and proteome of *C. tanaceti* was compared to genomes of related taxa using genome alignment, synteny and orthology analyses as following.

Genome alignment and synteny analysis. *Colletotrichum tanaceti* genome contigs were aligned to 13 other publicly available genomes (Table 1) of *Colletotrichum* species using Nucmer in Mummer v 4.0 [78, 79]. Contig-alignments were then filtered for a minimum 30% nucleotide identity and 200 bp in aligned length. The global coverage of each of the genomes from contigs of *C. tanaceti* was computed as a measure of pairwise sequence comparison between the two genomes.

The program 'Synteny Mapping and Analysis Program', SyMAP v 4.2 [80, 81] was used to map *C. tanaceti* contigs of the highest sequence length (>150 kb) to the chromosomes of *C. higginsianum* IMI349063 reference genome [43] to identify the syntenic regions. PROmer was invoked within SyMAP.

Table 1. Genomes used in the comparative genomic analyses.

Organism	Identifier[a]	Taxonomy ID[b]	Genbank accession number[c]	Bio project ID[d]	Strain[e]	Assembly version	Reference
Colletotrichum chlorophyti	CCh	708187	MPGH00000000.1	PRJNA350752	NTL11	ASM193710v1	[82]
Colletotrichum fioriniae	CFi	1445577	JARH00000000.1	PRJNA233987	PJ7	GCA_000582985.1	[83]
Colletotrichum fructicola	CFr	1213859	ANPB00000000.1	PRJNA225509	Nara gc5	GCA_000319635.1	[6]
Colletotrichum gloeosporioides	CGl	1237896	AMYD00000000.1	PRJNA176412	Cg-14	GCA_00446055.1	[84]
Colletotrichum graminicola	CGr	645133	ACOD00000000.1	PRJNA37879	M1.001	M1_0001_v1	[43]
Colletotrichum higginsinum	CHi	759273	LTAN00000000.1	PRJNA47061	IMI 349063	GCA 001672515.1	[43]
Colletotrichum incanum	CIn	1573173	LFIW00000000.1	PRJNA286717	MAFF 238704	GCA_001189835.1	[9]
Colletotrichum nymphaeae	CNy	1460502	JEMN00000000.1	PRJNA237763	IMI 504889	GCA_001563115.1	[7]
Colletotrichum orchidophilum	COc	1209926	MJBS00000000.1	PRJNA411788	IMI 309357	GCF_001831195.1	[85]
Colletotrichum orbiculare	COr	1213857	AMCV00000000.1	PRJNA171217	MAFF 240422	Corbiculare240422v01	[6]
Colletotrichum salicis	CSa	1209931	JFFI00000000.1	PRJNA238477	CBS 607.94	GCA_001563125.1	[7]
Colletotrichum simmondsii	CSi	703756	JFBX00000000.1	PRJNA239224	CBS 122122	GCA_001563135	[7]
Colletotrichum sublineola	CSu	1173701	JMSE00000000.1	PRJNA246670	TX430BB	GCA_000696135.1	[86]
Colletotrichum tanaceti	CT1	1306861	PJEX00000000	PRJNA421029	BRIP57314		
Verticillium dahliae	VDh	498257	ABJE00000000.1	PRJNA225532	VdLs.17	GCF_000150675.1	[87]
Botrytis cinerea	BCi	332648	AAID00000000.2	PRJNA15632	B05.10	GCF_000143535.2	[88]
Sordaria macrospora	SMa	771870	CABT00000000.2	PRJNA51569	k-hell	GCF_000182805.2	[89]
Fusarium oxysporum	FOx	426428	AAXH00000000.1	PRJNA18813	CBS 123668	GCF_00149955.1	[90]

[a] Short identifier used in pace of the species name in supplementary information

[b] Taxonomy ID of each species according to the NCBI taxonomy database

[c] Genbank accession number for the deposited nucleotide sequence

[d] NCBI bioproject ID

[e] version of the genome assembly

https://doi.org/10.1371/journal.pone.0212248.t001

Orthology search and phylogenomics analysis. The proteomes of *C. tanaceti* and the publicly available 17 other species were subjected to ortholog searching using OrthoMCL v2.0.9 [71] and MCL [91] with an inflation value of 1.5. The orthoMCL output was used to determine the percent orthology among the species and to determine the core gene set for *Colletotrichum*. The ortho-groups with pathogenicity genes (inferred as below) of *C. tanaceti* were extracted and used to determine the percent conservation of those gene categories within the genus. Furthermore the single copy orthologs were extracted from the orthoMCL output and aligned using MAFFT v.7 [92]. These alignments were then trimmed using trimAl v.1.3 [93] to remove all positions in the alignment with gaps in 20% or more of the sequences, unless this leaves less than 60% of the sequence remaining. The trimmed reads were concatenated using FASconCAT-G [94]. The concatenated alignment was partitioned and amino acid substitution models were predicted for each partition using ProtTest 3 [95] in FASconCAT-G. The partitioned, concatenated alignment was subjected to maximum likelihood phylogenetic analysis using RAxML v8.2.10 [96] to find the best tree from 20 maximum likelihood searches and using 100 bootstrap replicates. Evolutionary distance in number of substitutions per site was computed using the *ape* package [97] in the R statistical language framework v 3.5.1. [98] from the maximum likelihood tree.

Estimation of divergence dates. The phylogram developed from above was utilized to estimate the divergence dates of the species considered as following. The final RAxML phylogenetic tree was used to generate an ultrametric tree in r8s v1.81 [99] applying the penalized likelihood method [100] and the truncated Newton (TN) algorithm [101]. Divergence times were estimated using previously derived estimates [8, 11, 102] of 267–430 million years (Myr) for the Leotiomycetes-Sordariomycetes crown, 207–339 Myr for the Sordariomycete crown and 45–75 Myr for the *Colletotrichum* crown as calibrations. An optimal smoothing factor which was deduced using the cross validation process [99] among 50 values across 1 to 6.3e +09 was used in the divergence time estimation.

Prediction of secretome and database searches for identifying other virulence factors

Predicted proteins of *C. tanaceti* were used in downstream prediction of the secretome [103]. A union of three software tools: SignalPv4.1 [104], Phobius [105] and WoLFPSORT [106] was used to predict the candidate proteins to be used downstream of the pipeline. Proteins with either signal peptides predicted using SignalP or Phobious or proteins predicted as 'extracellular' in WoLFPSORT were retained as candidates for secreted proteins. Proteins with transmembrane domains were identified using TMHMM v.2.0 [107] and were excluded as secreted proteins. Proteins with signals targeting the endoplasmic reticulum and GPI anchors were identified and excluded using Ps-SCAN [108] and Pred-GPI [109] respectively. NLStradamus [110] was used to identify proteins with nuclear localization signals. Curated secretome was subjected to homology search against the CDD database to identify the conserved domains (E-value $\leq$ 1e-10). The candidate secreted effector proteins were identified by passing the secretome through the program EffectorP v1.0 [21]. Predicted effector candidates were manually inspected and candidates with known plant cell wall degrading catalytic domains, such as cutinases (PF01083.21), short-chain dehydrogenases (PF00106.24), glycosyl hydrolases (PF00457), peptidases (PF04117.11) and lipases (PF13472.5) were excluded. Candidates with no detectable conserved domains and no homology (E-value $\leq$ 1e-3) to any other proteins in NCBI–nonredundant protein sequence database were defined as species-specific. Putative secreted peptidases and inhibitors were predicted by stand-alone blastp (E-value $\leq$ 1e-10) homology searches of the MEROPS-MPEP database (consisting only the sequences of peptidase and

inhibitor units) of MEROPS release 12.0 [111]. Furthermore, potential virulence factors of *C. tanaceti* were identified by blastp searches (E-value ≤ 1e-10) against PHI-base v 4.4 [28]. The online analysis tools, Antibiotics and Secondary Metabolite Analysis Shell (antiSMASHV.4) [30] with default parameters and SMURF [112] were used to predict potential secondary metabolite backbone genes and clusters using the default parameters. Cytochrome P450s and transporters were described based on blastp (E-value ≤ 1e-10) homology searches against the Fungal Cytochrome P450 database [113] and the Transported Classification Database [114]. The functional annotations for *C. tanaceti* were compared across 17 other closely related taxa (Table1). The family specific Hidden Markov Model profiles of dbCan database v6 [115] were employed using the program HMMScan in HMMER v31.b2 [116] in order to identify the carbohydrate active enzymes (CAZymes) and the CAZyme families in the proteome of *C. tanaceti*. Fungi-optimal cut-off E-value of 1e-17 and a coverage cut-off of 0.45 [115] were used in the analysis which was repeated for seventeen related species (Table 1). The identified CAZymes were run though InterProScan 5 [73] (E-value ≤ 1e-10) to check for false positives. The member counts of each CAZyme family for each taxon were corrected accordingly.

Evolution of CAZyme gene families

CAFE v4.0 [117, 118] was used to estimate the number of CAZyme gene family expansions, contractions and the number of rapidly evolving gene families upon divergence of different lineages. Error-models [118] were estimated to account for the genome assembly errors and were incorporated into computations. A universal lambda value (maximum likelihood value of the birth-death parameter) was assumed and gene families with significant size variance were identified using a probability value cut-off of 0.01. The branches responsible for significant evolution, were further identified using the Viterbi algorithm [117] with a probability value cutoff of 0.05. Sizes of plant pathogenicity-related gene families from CAZomes of each of the species; the 'CAZyme pathogenicity profiles' were retrieved and compared using the online tool ClustVis [119]. The 'CAZyme pathogenicity profile' of a particular species included the gene families that have activities in binding to or degradation of plant cell wall components such as cellulose, hemicelluloses, lignin, pectin, cutin and chitin.

Testing for the bipartite nature of the *C. tanaceti* genome

The GC-bias of the genome was detected using OculterCut version 1.1 with default settings [37]. The genome wide dinucleotide frequency and the two RIP indices, TpA/ApT and (CpA + TpG)/(ApC + GpT) were computed and the RIP affected genomic regions were identified using the RIPCAL V.2 [120]. Significant enrichment of A-T rich regions of the *C. tanaceti* genome with interspersed repeats and RIP were tested using permutation tests implemented in the package regioneR in the R statistical language framework v3.5.1 [98] with the evaluation function for number of overlaps [121]. Ten thousand random iterations were conducted, from which a *Z*-statistic estimate, and its associated probability, were computed.

Relationship of pathogenicity related genes with repeat elements and RIP

The mean distances between putative pathogenicity genes and 1) repeats, 2) RIP affected regions were analyzed using permutation tests implemented in the package regioneR [121] R statistical language framework v3.5.1 [98]. Repetitive element categories incorporated in this analysis included: 1) tandem and interspersed repeats combined; 2) tandem repeats; and 3) interspersed repeats. These were compared to the pathogenicity related gene classes: 1) CAZymes; 2) peptidases; 3) secondary metabolite biosynthetic gene clusters; and 4) effectors. The mean distance between each gene in above categories and the nearest repetitive element/

RIP affected region was compared against a distribution of distances of random samples from the whole genome. Ten thousand random iterations were conducted, from which a *Z*-statistic estimate, and its associated probability, were computed for each gene category.

Results

Colletotrichum tanaceti genome and gene content

The genome of isolate BRIP57314 was assembled into 5,242 contigs with an N50 value of 103,135 bp and assembly size of 57.91Mb. The average GC content was 49.3% (Table 2). The genome size and GC content of *C. tanaceti* was within the range previously reported for other *Colletotrichum* spp. (S1 Fig). Draft genome assembly and the raw unassembled sequences are available under the accession no PJEX00000000 in Genbank. The genome contained 12,172 coding genes with an average gene length of 2,575bp. Mean exon count per gene was 3, and 54.1% of the genome sequence contained protein-encoding genes. In the BUSCO analysis, out of the 3,725 benchmarking genes in the Sordariomycetes group, the genome was reported to contain 3,656 complete BUSCOs (98.2%), of which two were duplicated and the rest were single copy genes (98.1%). A total of 30 (0.8%) BUSCOs were fragmented and 39 were missing (1.0%). The repeat content of *C. tanaceti* was 24.6% of the total genome of which 85.2% was interspersed repeats (Table 3).

Of the 12,172 predicted proteins, 11,352 had an annotation edit distance (AED) value of less than 1.0, and 2962 genes had an AED value of zero. The number of genes without putative annotation from the public database searches was only 958. A total of 8,945 proteins (73.5% of proteome) had InterProScan annotations of which 6,911 contained 9,647 Pfam domain annotations and 5,452 had GO term ontology annotation. The most abundant (n = 129) Pfam domain was the cytochrome P450 family (PF00067) followed by the protein kinase domain

Table 2. Features of the *Colletotrichum tanaceti* BRIP57314 genome.

Feature	Statistics
GC content (%)	49.3
N50 (bp)	103,135
Maximum sequence length (bp)	945,015
Mean length (bp)	11,047
Number of base pairs	57,912,474
Number of contigs	5,242
Number of genes	12,172
Number of exons	35,792
Number of introns	23,620
Number of CDS	12,172
Overlapping genes	3,983
Contained genes	1,586
Mean gene length (bp)	2,575
Mean exon length (bp)	787
Mean intron length (bp)	137
Mean CDS length (bp)	1,440
% of genome covered by genes	54.1
% of genome covered by CDS	30.3
Mean mRNAs per gene	1
Mean exons per mRNA	3
Mean introns per mRNA	2

https://doi.org/10.1371/journal.pone.0212248.t002

Table 3. Repetitive elements of the *C. tanaceti* genome.

Repetitive element	Number of elements	Length occupied (bp)	Percentage of sequence
SINEs:	49	4,123	0.01
ALUs	0	0	0
MIRs	11	869	0
LINEs:	612	251,619	0.43
LINE1	207	48,554	0.08
LINE2	35	2,588	0
L3/CR1	82	5,928	0.01
LTR elements:	7,299	4,825,086	8.33
ERVL	2	120	0
ERVL-MaLRs	1	39	0
ERV_classI	3	209	0
ERV_classII	1	32	0
DNA elements:	1,436	905,846	1.56
hAT-Charlie	3	140	0
TcMar-Tigger	6	529	0
Unclassified:	8,863	6,153,241	10.62
Total interspersed repeats		12,139,915	20.96
Small RNA:	754	210,370	0.36
Satellites:	0	0	0
Simple repeats:	9,941	1,757,918	3.04
Low complexity:	3,064	147,883	0.26
Total repeat content			24 .62%

https://doi.org/10.1371/journal.pone.0212248.t003

(*n* = 127; PF00069). Gene enrichment analysis suggested enrichment of many GO terms including those associated with translation and chromosome telomeric region (S1 Table). Putative proteins of *C. tanaceti* were subjected to KEGG pathway analysis which returned assignment of 5,883 proteins to known pathways (S2 Table). The highest number of KO identifiers was among the metabolic pathway assignments (*n* = 693) of which the majority (*n* = 363) were for amino acid metabolism followed by carbohydrate metabolism (*n* = 290) (S3 Table). Among the environmental information processing pathways, 81 *C. tanaceti* genes were assigned into 47 KO identifiers belonging to MAPK pathway (S4 Table). Furthermore, 24 *C. tanaceti* proteins were annotated with 10 aflatoxin biosynthesis pathway KO assignments (S5 Table) and 56 proteins were assigned KOs for ABC transporters (S6 Table).

Genome alignment and synteny

The global alignment coverage of 13 other *Colletotrichum* genomes from *C. tanaceti* contigs was proportionate to the evolutionary proximity to *C. tanaceti* (Fig 1A). The highest coverage was in *C. higginsianum* (63.8%) and the least was in *C. orbiculare* 4.26%. Among the *C. tanaceti* contigs aligned to the chromosomes of *C. higginsianum*, the best alignment coverage was to chromosome NC_030961.1 (chromosome 9) (S7 Table). *Colletotrichum tanaceti* contigs (*n* = 155 of size≥10 kb) were mapped in SyMAP synteny analysis to form 142 synteny blocks which covered 44.0% of the *C. higginsianum* and 80.0% of the *C. tanaceti* sequences that were used (S2 Fig). Genes were present in 92.0% of the syntenic regions in *C. tanaceti* and in 77.0% of *C. higginsianum*. No inverted synteny blocks were reported. Despite the highest coverage in *C. higginsianum* chromosome 9, the largest synteny block was identified between the complete *C. tanaceti* contig 4 (945.01 kb of length) and *C. higginsianum* chromosome NC_030954

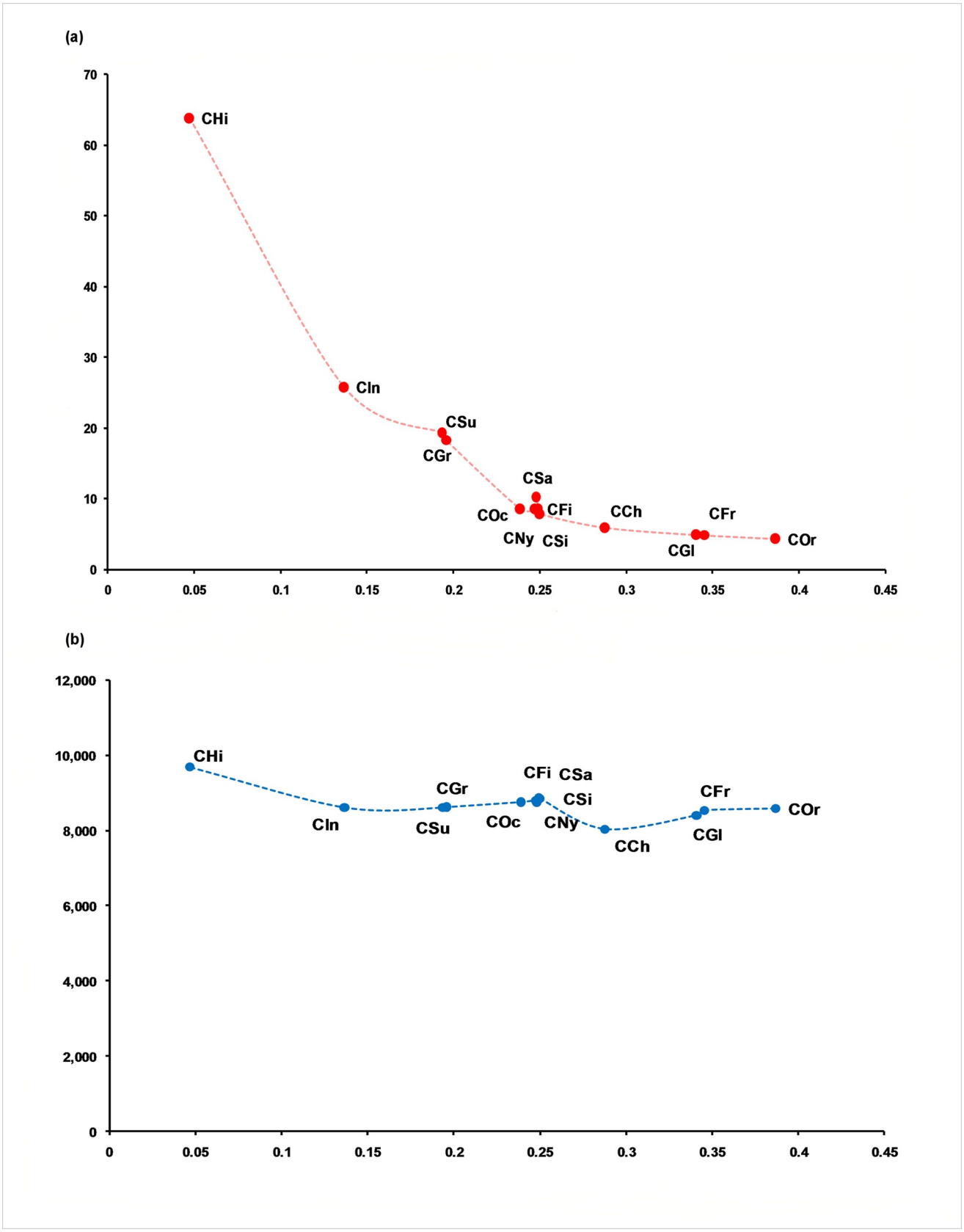
(a)
70
60
50
40
30
20
10
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
CHi
CIn
CSu
CGr
CSa
COc
CFi
CCh
CFr
CNy
CSi
CGl
COr
(b)
12,000
10,000
8,000
6,000
4,000
2,000
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
CHi
CIn
CGr
CSu
CFi
CSa
COc
CSi
CNy
CCh
CFr
CGl
COr

Fig 1. Comparison of the *C. tanaceti* genome to previously published *Colletotrichum* spp. genomes. (a) Percentage global alignment (y axis) of 13 *Colletotrichum* draft genomes to contigs representing the *C. tanaceti* draft genome, plotted against evolutionary distance with reference to *C. tanaceti* (x axis), (b) Number of orthologs shared by 13 *Colletotrichum* draft genomes and *C. tanaceti* (y axis) plotted against the evolutionary distance with reference to *C. tanaceti* (x axis); evolutionary distance given in number of substitutions per site, computed using the ape package [98] in R from a maximum likelihood tree.

https://doi.org/10.1371/journal.pone.0212248.g001

(Chromosome 1). A total of 38 effector candidates of *C. tanaceti* were within these syntenic regions between *C. tanaceti* and *C. higginsianum*. No synteny blocks were detected to the two mini chromosomes (NC_030963.1 and NC_030964.1) of *C. higginsianum*.

Orthology search

Of 221,456 total genes from 18 genomes, the number of core genes reported for all ascomycetes in the orthology analysis was 3,944. A total of 10,695 putative proteins from *C. tanaceti* were assigned to 10,074 groups containing orthologs and/or recent paralogs and/or co-orthologs across all species tested. A total of 6,002 genes were conserved in all tested members of the genus *Colletotrichum*. *Colletotrichum tanaceti* had 9,679 orthologs with *C. higginsianum* which was the highest ortholog count among *Colletotrichum* spp. followed by 8,855 orthologs with *C. nymphaea* (Fig 1B). Twenty of these groups, with 48 genes among them were exclusive to *C. tanaceti* and were defined as recent paralogs (*in-paralogs*) of *C. tanaceti* with no homology to the 16 other species tested.

Divergence time in *Colletotrichum* lineages

A total of 2,214 single copy ortholog (SCO) genes identified among the *C. tanaceti* and 17 closely related genomes (Table 1) were used to generate a maximum likelihood (ML) evolutionary tree in which all branches achieved bootstrap support of 100%. *Colletotrichum tanaceti* formed a clade with *C. higginsianum*, a member of the destructivum complex and the two destructivum complex members formed a sister clade with the graminicola complex members and *C. incanum*. A smoothing factor value of 1 was reported as the optimal value for divergence time predictions in r8s. *Colletotrichum tanaceti* and *C. higginsianum* were reported to have diverged ~9.97 million years ago (mya). The most recent common ancestor (MRCA) of gloeosporioides, acutatum, and graminicola clades were reported to be 6.12, 10.98 and 15.78 mya, respectively (Fig 2).

Identification of pathogenicity related genes in *C. tanaceti*

Secretome of *C. tanaceti*. Of the 12,172 putative proteins, 1,024 (8.41%) were predicted to be secreted. A total of 2,702 Conserved Domain Database (CDD) domains were found in the secretome. Of these, 287 were specific features with NCBI curated models, 124 were generic features with only the superfamily annotations [122]. Only 433 queries had no known domain hits. The secretome was rich in alpha beta hydrolase superfamily (cl21494) containing enzymes, glycosyl hydrolases and proteolytic enzymes and cytochrome P450 monoxygenases (*P450*) (S8 Table). A total of 100 secreted proteins had nuclear-localization signals (S8 Table).

A total of 233 putative effector candidates were predicted by EffectorP. Following manual inspection and filtering out candidates with known plant cell wall degrading catalytic domains, a total of 170 candidates were selected as putative effectors of *C. tanaceti* (S9 Table). The putative secreted candidate effector repertoire of *C. tanaceti* contained homologs of known effectors, such as the *Ecp6* of *Cladosporium fulvum* [123], *MC69* of *Magnoporthe oryzae* and *Colletotrichum orbiculare* [124], *EP1* of *Colletotrichum graminicola* [125], *NLP1* of *Colletotrichum higginsianum* [126] *ToxB* of *Pyrenophora tritici repentis* [127] and *Magnoporthe oryzae*

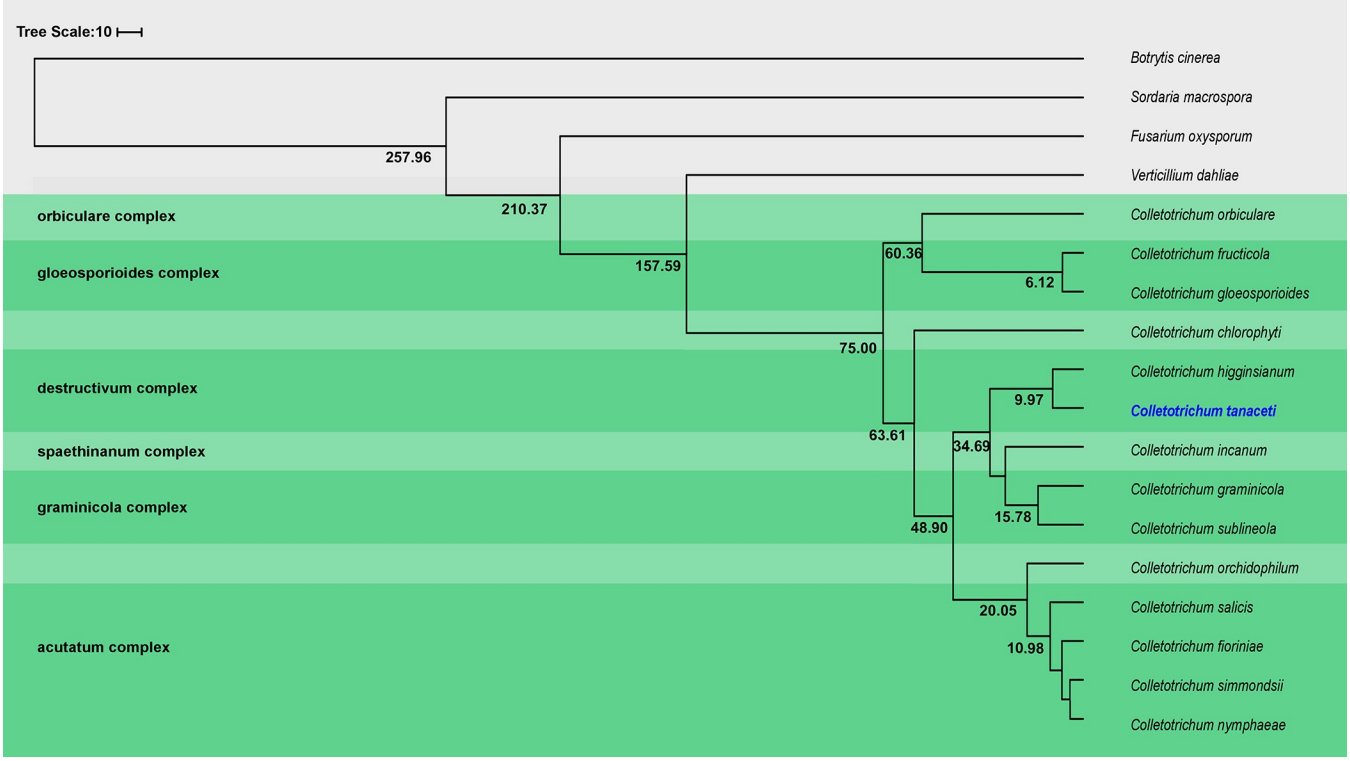

Fig 2. Chronogram showing divergence time estimations (in million years) for *Colletotrichum* spp. and related taxa.

https://doi.org/10.1371/journal.pone.0212248.g002

Bas3 [128]. Furthermore, among the putative effector candidates, there were proteins with conserved domains of known virulence factors. Most effector candidates were small (average length of 157 amino acids) and rich in cysteine (average cysteine composition was 3.3%) which are the hallmarks of effectors. A total of 78 conserved motifs of fungal effectors [129] were present in 62 effector candidates which had at least one motif each. Twenty-two effector candidates that did not cluster in ortholog search among the 14 *Colletotrichum* and three related species, and also did not show detectable homology to the NCBI-nr and swissprot databases were defined as *C. tanaceti*-specific. Only 25% of the putative effectors of *C. tanaceti* were conserved among all 14 *Colletotrichum* spp.

A total of 98 putative secreted peptidases were predicted with the majority ($n = 64$) being serine peptidases largely comprising the S08 and S09 subfamilies. The second most abundant class was the metallo peptidases ($n = 19$) (S10 Table). All six putative aspartic peptidases belonged to subfamily A01. A total of 20 putative secreted peptidase inhibitors were reported in *C. tanaceti* comprising two carboxypeptide-y inhibitors, five family-19 inhibitors and 13 family-14 inhibitors (S11 Table). Forty nine percent of the putative proteases of *C. tanaceti* were among the "core" set of proteases of *Colletotrichum* spp. tested.

Secondary metabolite-related genes and clusters. Forty-one putative secondary metabolite backbone genes were predicted in *C. tanaceti* using SMURF and the majority were polyketide synthases (PKS, $n = 13$) with four PKS-like proteins. Furthermore, nine non-ribosomal peptide synthases (NRPS), eight NRPS–like proteins, two hybrid PKS-NRPS enzymes and five dimethylallytryptophan synthases (DMATS) were also predicted as backbone genes (S12 Table). A total of 52% of these putative backbone genes were within the core set of genes in *Colletotrichum*. A total of 33 putative secondary metabolite gene-clusters were predicted

surrounding the backbone genes. However, the program antiSMASH predicted a total of 50 clusters. Among the clusters, there were twelve typeIPKS, two typeIIIPKS, thirteen terpenes, eleven NRPS, four indoles, three typeIPKSs-NRPS, one typeIPKS-indole and four other proteins. Cluster 10 of typeIPKS showed 100% similarity to the genes in LL-Z1272 beta biosynthetic gene cluster (BGC0001390_cl). Furthermore, a homolog to the melanin biosynthetic gene *SCD1* was also reported in *C. tanaceti* (CTA1_6632). When predictions from the two tools were compared, putative SMB clusters on 31 contigs of *C. tanaceti* were predicted by both tools and 19 of the backbone genes from SMURF were also predicted in antiSMASH (S12 Table). A total of 37 putative SM clusters were within the syntenic blocks of *C. higginsianum*. The conserved SM domains identified in each cluster were reported (S13 Table). Predictions from antiSMASH were compared across taxa and majority of the clusters were typeIPKS like followed by NRPS in all ascomycetes compared (Fig 3A). The highest number of clusters were reported from *C. fructicola* (n = 84) followed by *C. higginsianum* (n = 74) and *C. gloeosporioides* (n = 73). The composition of the SMB gene cluster composition of *C. tanaceti* was most similar to *C. orchidophilum*, the acutatum complex members and *C. orbiculare* (S3 Fig).

Cytochrome P450 monoxygenases (P450s) and transporters. In the *C. tanaceti* genome, 1,457 putative genes had homologs in the fungal cytochrome P450 database (S14 Table) and 911 out of that had >30% identity. There were 1,824 homologs (S15 Table) in the transport classification database for *C. tanaceti* with 1,276 genes with >30% identity. The majority (n = 430) of the homologs were genes of the major facilitator superfamily (MFS, 2.A.1) followed by 129 genes of the ABC transporter family (3.A.1) and 123 of N.P.C 1.I.1. Within *Colletotrichum* genus, members of the gloeosporioides complex had the highest number of homologs for both P450s and transporters (Fig 3B).

Homologs in PHI-base. A total of 3,497 homologs were recorded in *C. tanaceti* from the pathogen-host interaction database (PHI), of which 1,592 represented homologs of genes that result in reduced virulence in loss of function mutants (S16 Table). The second most common (n = 1,514) were the unaffected pathogenicity category, 382 homologs were for loss of pathogenicity and 42 were in the effector category. Notably, 141 homologs were reported to genes of which the loss of function mutants were lethal to the pathogen and 103 homologs were reported to genes in which virulence increased after loss of function mutation (Fig 3C). The two gloeosporioides complex members had the highest number of homologs in the database among the *Colletotrichum* spp., followed by the acutatum complex species, *C. simmondsii*, *C. fioriniae* and *C. nymphaea*. Despite *C. higginsianum* having a large number of homologs, *C. tanaceti* had a below average number for all the categories among the *Colletotrichum* spp., with a profile similar to *C. orchidophilum*, *C. chlorophyti* and *C. graminicola* (S4 Fig).

CAZymes. A total of 608 *C. tanaceti* proteins were assigned to 121 CAZyme families of which 43% was putative glycosyl hydrolases followed by 18% of putative redox enzymes (auxiliary activities) and 14% putative carbohydrate esterases (S17 Table). Putative carbohydrate binding molecules and polysaccharide lyases both formed 7% each of the *C. tanaceti* CAZome whereas 11% was glycosyltransferases. A total of 179 CAzymes were secreted (S17 Table). Members of the gloeosporioides and acutatum complexes had the largest CAZomes among *Colletotrichum* spp. The CAZyme repertoires of the graminicola complex members were relatively small (Fig 3D).

Evolution of CAZyme families upon divergence of Colletotrichum lineages

A total of 152 CAZyme families, predicted at the node of MRCA for *S. macrospora* and *B. cinerea*, were used in gene family evolution analyses in CAFÉ. A uniform birth-death parameter (λ) of 0.0023 was computed. Thirty gene families were reported to be significantly evolving

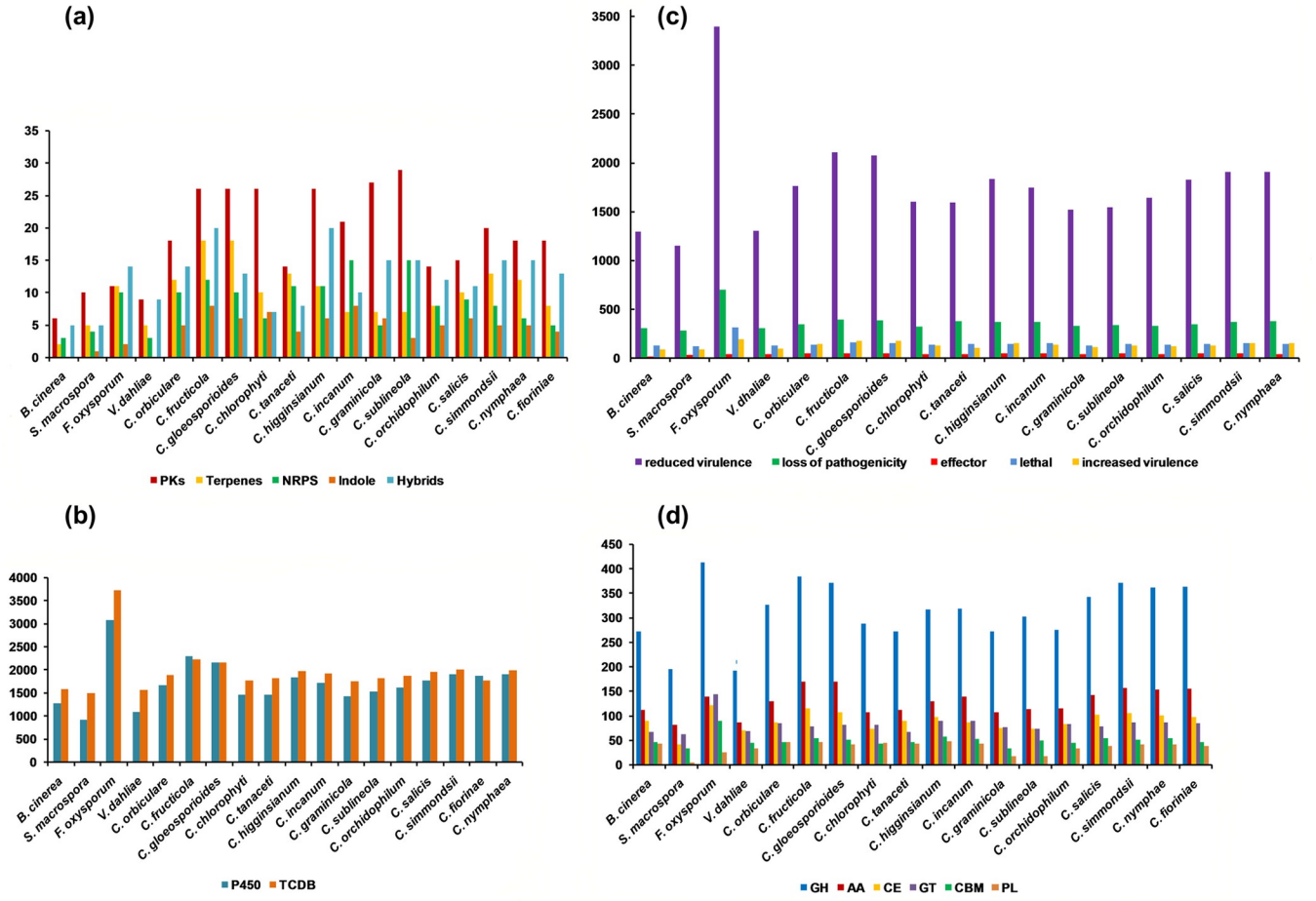

Fig 3. Composition of different pathogenicity gene categories predicted for *Colletotrichum tanaceti* and related species. The number of genes in each gene category (x axis) plotted for each species (y axis). (a) secondary metabolite biosynthetic gene clusters-(gene clusters producing polyketides, terpenes, non-ribosomal peptides (NRPS), indoles and the hybrids of above); (b) number of homologs in the fungal cytochrome P450 database and the transporter classification database (TCDB); (c) homologs in the pathogen-host interaction database; homologs to entries in the "unaffected pathogenicity" database were excluded; (d) CAZyme classes; glycoside hydrolases (GH), polysaccharide lyases (PL), glycosyltransferases (GT), carbohydrate esterases (CE), molecules with auxiliary activities (AA), and carbohydrate binding molecules (CBM).

https://doi.org/10.1371/journal.pone.0212248.g003

(family-wide p value ≥ 0.05), of which 21 were rapidly evolving (family-wide $p \geq 0.01$ and *Viterbi* $p \geq 0.01$ in any lineage) (S18 Table).

At the divergence of *Colletotrichum* spp., 39 expansions and 12 contractions were predicted with respect to its MRCA with *Verticillium* species (S19 Table). Expansions included the lignin hydrolase family AA2, pectin degrading polysaccharide lyase families (PL1, 3, 4, 9 and GH78), lignocellulose degrading families (AA3, AA9, GH131, GH5, GH6, GH7), hemicelluloses degrading families (CE1, CE4, CE5, CE12, GH3, GH16, GH30, GH43, GH51, GH67, and GH10), Lys M domain containing family CBM50 and cutinase family CE5. The cellulose degrading family GH131 was the only rapidly evolving CAZyme family (family-wide $p \geq 0.01$ and *Viterbi* $p \geq 0.01$) which expanded upon the divergence of *Colletotrichum* spp. Within the genus, the highest number of expansions ($n = 38$) was reported at the divergence of the gloeosporioides-complex clade with only 4 contractions. Notably, the CBM18 and GH10 families were contracted and many families with plant cell wall degrading enzyme activity were expanded. The rapidly and significantly expanded families, (family-wide $p \geq 0.01$ and *Viterbi*

$p \geq 0.01$) upon the divergence of the gloeosporioides-complex clade include GH43, GH106, CBM50 and AA7. At the divergence of the acutatum-complex clade, there were 22 expansions, of which expansions in GH78, GH43 families were rapid and significant and there was only one contraction. The divergence event of the graminicola-complex clade involved contractions in many CAZyme families with pectin degradation activity showing significant, rapid contractions (family-wide $p \geq 0.01$ and *Viterbi* $p \geq 0.01$)in families AA7, CBM50, CE8, GH28, GH78, PL1, and PL3. Divergence of the destructivum complex-clade was associated with 11 expansions and 21 contractions, of which expansion in AA7, GH74 and CE10 was significant and rapid.

Among the other species considered, *Fusarium oxysporum* had the highest number of genes ($n = 344$) that were gained, with 75 expanded CAZyme with respect to its MRCA (S20 Table). *Colletotrichum incanum* had the second highest number of gene family expansions ($n = 35$) and genes gained ($n = 69$) followed by *C. higginsianum* (31 and 68 respectively). Forty CAZyme families contracted and only nine expanded in *C. tanaceti* with respect to the MRCA with *C. higginsianum*. The AA2 family with lignin peroxidase activity and the hemicellulose degrading GH12, GH74 families were among the expanded families, but many families with pathogenicity and plant cell wall degrading activity had contracted in *C. tanaceti*. However, the highest number of significant, rapidly evolving gene families was reported from *C. tanaceti* ($n = 9$) followed by *F. oxysporum* and *C. higginsianum*, both which had seven rapidly evolving gene families each. In *C. tanaceti*, rapidly evolving CAZyme families included AA9, GH131 with lignocellulose degrading activity, chitin binding molecule families CBM18 and CBM50, GH18 with chitinase activity, GH3 and GH74 with hemicelluloses degrading activity, GH78 with pectinase activity and GT1 with glucuronosyltransferase activity. However, CBM18 and GH74 were the only families that expanded among those above with the rest contracting in *C. tanaceti* with respective to their MRCA. Gloeosporioides complex species had the largest 'CAZyme pathogenicity profiles' among all *Colletotrichum* species considered. The CAZyme pathogenicity profile of *C. tanaceti* was most similar to those of *Colletotrichum* species known to have an intermediate host range, infecting many hosts within a single plant family or few hosts across several plant families (Fig 4). When compared the overall pathogenicity gene profiles of all *Colletotrichum* spp., which included the numbers of the SMB clusters, transporters, P450s, CAZymes and the homologs to the PHI database, the profile of *C. tanaceti* was most similar to *C. orchidophilum* and *C. chlorophyti* (Fig 5).

RIP affected regions of the *C. tanaceti* genome

The RIP indices computed for the genome of *C. tanaceti* using the dinucleotide frequencies (S21 Table) indicated strong RIP signals in the genome. The TpA/ApT index of *C. tanaceti* (1.2) was higher than the cutoff 0.89 and the (CpA + TpG)/(ApC + GpT) index (0.96) was lower than the cutoff 1.03 indicating a strong RIP signal. Homologs to two genes involved in RIP of *Neurospora crassa RID* [130] and *Dim-2* [131] were identified from the genome of *C. tanaceti* (CTA1_356s and CTA1_4791s respectively).

Bipartite nature of *C. tanaceti* genome

Distinct A-T rich regions and G-C equilibrated regions were identified in the genome of *C. tanaceti* (Fig 6). A total of 24.3% of the genome which had an average length of 3.77 kb was rich in A-T and had a maximum G-C of 29%. The *Z-score* and the *P value* of the permutation tests for the random association of transposable elements with the A-T rich regions were 799.354 and ≤ 0.001 respectively. The *Z-score* and the *P value* of the permutation tests for the enrichment of RIP with the A-T rich regions were 165.001 and ≤ 0.001 respectively. For the

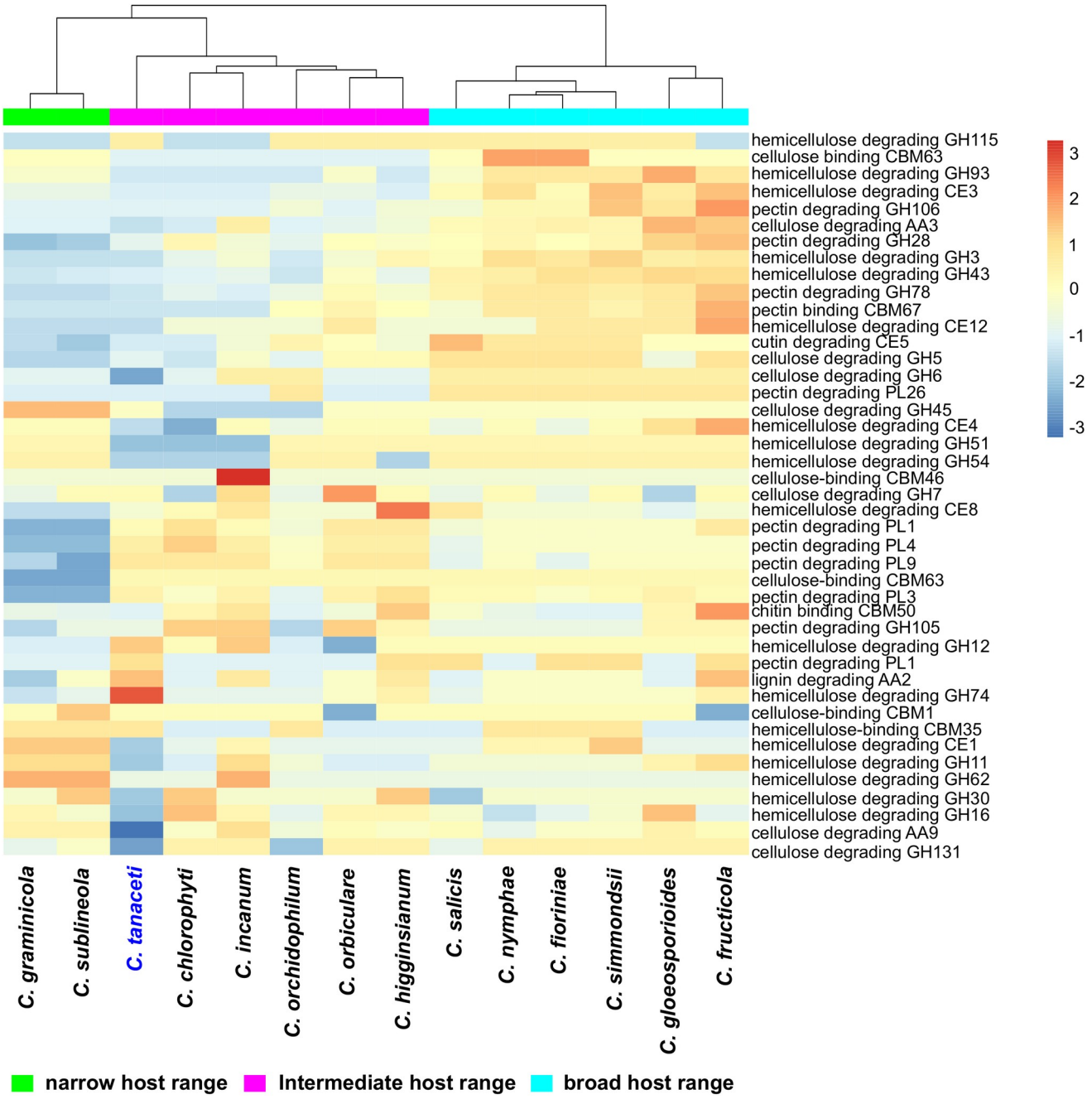

Fig 4. Comparison of CAZyme pathogenicity profiles predicted for *Colletotrichum* species. Number of genes in each CAZyme family is normalized using unit variance scaling. Hierarchical clustering performed with Euclidean distance and Ward linkage. Overrepresented and underrepresented CAZyme families are represented in red to orange and blue respectively as fold standard deviations above and below the mean.

https://doi.org/10.1371/journal.pone.0212248.g004

A-T rich regions of the genome, the TpA/ApT index was 1.86 and (CpA + TpG)/(ApC + GpT) index was 0.32 indicating a strong RIP signal. A total of 85 genes were reported in these regions which had a gene density of 6.04 genes per Mb but the majority (68.25%) of these genes was hypothetical. Two secondary metabolite biosynthetic genes, 3 CAZymes, 2 cytochrome P450s, 2 lipases, 4 transporters, one transcription factor and one DNA polymerase were putative pathogenicity genes among the genes in the A-T rich regions (S22 Table). The

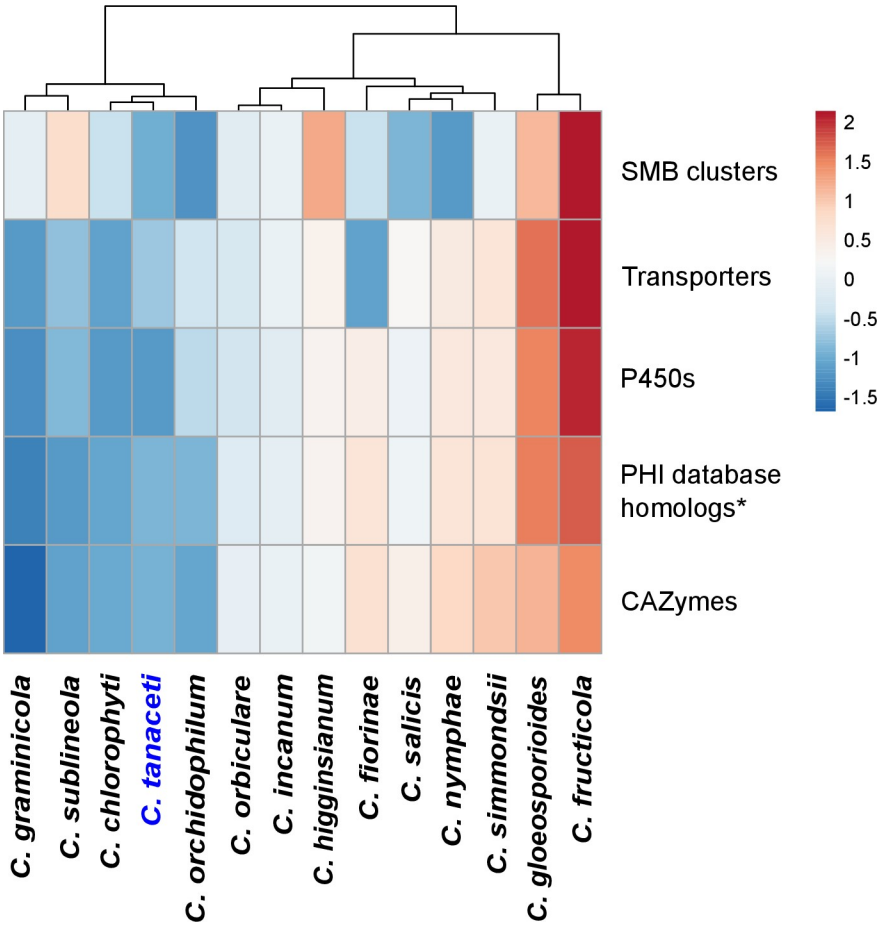

Fig 5. Comparison of the overall pathogenicity profiles predicted for *Colletotrichum* species. The numbers of CAZymes, secondary metabolite biosynthetic gene clusters (SMB), homologs in the transporter classification database (transporters), homologs in the fungal cytochrome P450 database (P450) and the number of homologs in the PHI database, excluding the homologs to entries in the "unaffected pathogenicity" database were used in the analysis. Hierarchical clustering was performed using Euclidean distance and Ward linkage methods. The number of genes in each pathogenicity gene category is normalized using unit variance scaling. Overrepresented and underrepresented pathogenicity gene categories are represented in red to orange and blue respectively as fold standard deviations above and below the mean.

https://doi.org/10.1371/journal.pone.0212248.g005

G-C equilibrated regions accounted for 75.7% of the genome and the average length was 14.6 Kb. The maximum G-C percentage in these regions was 55.6 and 12,087 genes were reported with a gene density of 276 genes per Mb.

Relationship of putative pathogenicity genes with repeat elements and RIP

The permutation tests confirmed that genes in all the tested pathogenicity-related gene categories are located significantly closer to tandem repeats than expected in a random sample (Table 4). The negative *Z*-scores confirmed the mean distance between those genes and the nearest repetitive element was less than mean of a random sample of the genome. Furthermore, all gene categories except the CAZymes were located significantly closer to the interspersed repeats. However, the expanded and the contracted subgroups of the total CAZome were significantly associated with interspersed repeats (Table 4). All pathogenicity gene categories except contracted CAZymes and effectors were also located closer to the RIP affected regions of the genome than expected.

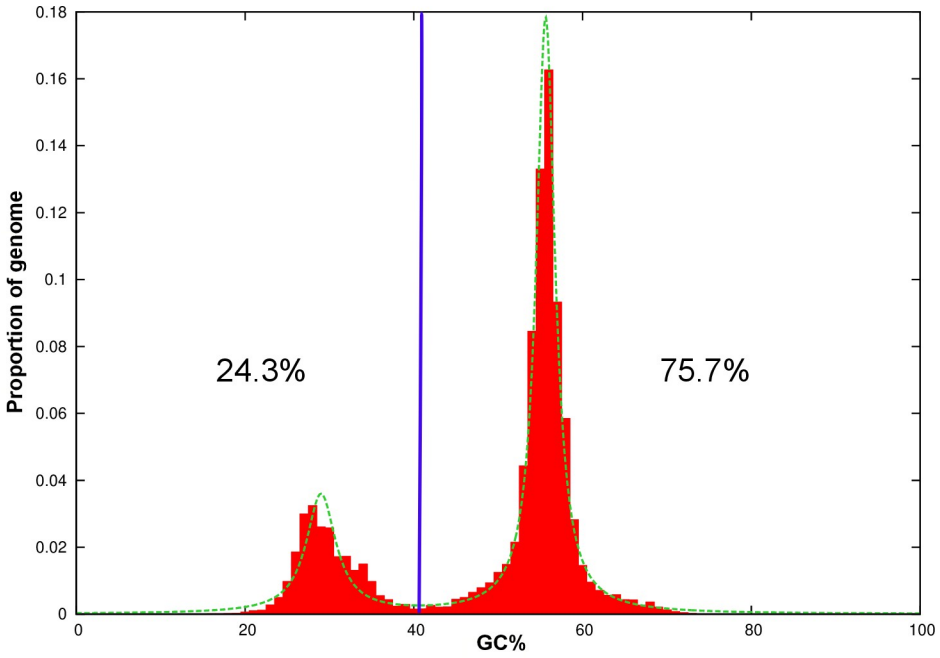

Fig 6. Plot of GC-content in the draft genome of *Colletotrichum tanaceti* against proportion of the genome. Genome segments were classified into A-T rich (24.3%) and G-C equilibrated (75.7%) using a GC content threshold of 40% (vertical blue line).

https://doi.org/10.1371/journal.pone.0212248.g006

Discussion

Genome and the repeat content of *Colletotrichum tanaceti*

This study reports the first draft genome sequence and annotations of the emerging plant pathogen, *C. tanaceti*. The high N50 value and BUSCO completeness indicates the high quality of the assembly and AED scores of less than one for the majority of predicted genes (93.3%) suggested that these genes had at least partial congruence with the transcriptomic evidence [132]. These good quality gene predictions and annotations will provide a solid foundation for downstream genetic, population genomic and evolutionary studies.

Table 4. Permutation tests for association of repetitive elements with pathogenicity gene categories.

Gene categories of interest	All repeats [a]		Tandem repeats		Interspersed repeats		RIP affected regions	
	Z score [b]	*P* value [c]	*Z* score [b]	*P* value [c]	*Z* score [b]	*P* value [c]	*Z* score [b]	*P* value [c]
CAZymes	-5.97	≤0.001	-3.914	≤0.001	-0.443	0.334	-4.674	<0.001
Expanded CAZymes	-3.514	≤0.001	-4.553	≤0.001	-3.050	≤0.001	-3.878	<0.001
Contracted CAZymes	-4.413	≤0.001	-3.237	≤0.001	-5.883	≤0.001	-1.534	0.06
Effectors	-5.631	≤0.001	-4.725	≤0.001	-3.861	≤0.001	1.3957	0.087
Peptidases	-5.787	≤0.001	-4.679	≤0.001	-3.895	≤0.001	-4.302	<0.001
SMB clusters	-7.901	≤0.001	-8.490	≤0.001	-2.610	0.003	-4.334	<0.001

[a] tandem and interspersed repeats

[b] Z-statistic estimate and its

[c] associated probability computed based on 10,000 random iterations.

https://doi.org/10.1371/journal.pone.0212248.t004

The genome of *C. tanaceti* had a larger repeat content (25%) than the typical 3–10% in fungi [133]. Simple sequence repeats comprised 3.03% of the genome of *C. tanaceti* which itself was unusually high for fungi (generally 0.08–0.67%) [134]. However, the majority of repeats were interspersed transposable elements (TE) (21%). TE content of *C. tanaceti* was higher than in six previously studied *Colletotrichum* species, including *C. higginsianum* which is in the same species complex, but lower than in *C. orbiculare* (44.8%). The majority of TE were retrotransposons, similar to other *Colletotrichum* spp. [41]. Proliferation of repetitive elements especially transposons, is known to be a major mechanism driving expansion of eukaryote genomes [135, 136]. Furthermore, TE activity favors chromosomal rearrangements, gene deletions, gene duplications and greater sequence diversity and is a mechanism of genome plasticity [35].

Colletotrichum tanaceti's genome consists of distinct A-T rich, gene sparse regions. These regions are also enriched in putative TE and RIP. Strong genome-wide RIP signals were observed on *C. tanaceti*. RIP is a method of controlling TE proliferation in fungi and facilitates genome plasticity, via high mutational rates and inactivating genes [35]. Homologs of two genes involved in RIP were present in *C. tanaceti* similar to *C. higginsianum* [15]. Therefore, TE proliferation in *C. tanaceti* may have caused accumulation of RIP as a control mechanism. These RIP mutations may have caused the bipartite nature of the *C. tanaceti* genome giving rise to A-T rich blocks [38–41] similar to the observations in *C. orbiculare* and *C. graminicola* [6, 41]. Similar to *C. orbiculare* and *C. graminicola* [6, 41], *C. tanaceti* has a known sexual stage [17] which could have activated the RIP in the genome [137]. Although genome-wide RIP was not prominent, the TEs in the genomes of *C. fructicola* [42], *C. higginsianum* [15], *C. truncatum* [41] and the *C. cereale* [138] has shown signals of RIP.

Pathogenicity genes of *C. tanaceti*

A large array of putative genes related to pathogenicity was inferred from the sequenced genome of *C. tanaceti*. Apart from many plant cell wall-degrading enzymes, effectors, *P450*s and the proteolytic enzymes, there were proteins with CFEM domain (pfam05730) [139] with roles in conidial production and stress tolerance [140] among the secreted proteins. The average cysteine composition, length and proportion of specificity of the candidate secreted effectors of *C. tanaceti* were similar to those hemibiotrophic pathogens [141]. However, a minority of effector candidates was neither small (<300bp) nor rich in cysteine (>3%), similar to previous reports of atypical effectors [142]. Effector candidates with a nuclear localization signal might translocate to the host nucleus and reprogram the transcription of genes related to host immune responses. Homologs to known effectors, and effectors with conserved domains of virulence factors may have similar functions in *C. tanaceti*, for example, in penetration peg formation (cyclophilin) [143], phytotoxity induction (cerato-platanin) [144] and adherence of the fungal structures to other organisms (hydrophobin) [145].

Most secreted proteases of *C. tanaceti* were serine proteases predicted to evade plant immune responses by degrading plant chitinases [22]. Subtilisins (S08) were the most abundant of these in *C. tanaceti*, similar to reports in other fungi [22]. Subtilisins, with their alkaline optima, and the proteases in other subfamilies with acidic optima, such as A01, C13, G01, M20 and S10 [146], might enable *C. tanaceti* to degrade plant proteins across a wide pH range. Also, the protease inhibitors of *C. tanaceti* might have effector-like roles via inhibition of plant defense proteases [147].

The SMB gene clusters and the candidate proteins of MAPKs pathways identified in the genome of *C. tanaceti* are also believed to play an important role in pathogenesis. The majority of the secondary metabolite clusters of *C. tanaceti* were typeI PKs-like which are usually

associated with synthesizing fungal toxins [148]. Melanin, another important secondary metabolite aids penetration via increasing turgor pressure [149]. Even though the gene cluster associated with melanin biosynthesis was not identified, the homolog of the melanin biosynthetic gene *SCD1* encoding Scytalone dehydratase [150] in *C. tanaceti* is worth investigating further since *SCD1* has been successfully used as a target for fungicides to control other pathogens [151]. Apart from their function in SM biosynthesis, the candidate P450s of *C. tanaceti* could be involved in housekeeping roles and therefore, could be good targets for fungicide development, as in the case of azoles targeting CYP51 [152]. Furthermore, the candidate proteins of MAPKs pathway in *C. tanaceti* could play a crucial role in appressorium formation [25, 153], penetration [154], conidiation [155] and pathogenesis-related morphogenesis [156], as reported for *C. higginsianum* and *C. lagenaria*.

Of the CAZyme families identified to be expanded in *C. tanaceti*, the chitin binding family CBM18 could play a role in protecting the *C. tanaceti* cell wall from exogenous chitinases, as is the case in *Trichoderma reesei* [157]. The expansion of the hemicellulose-degrading GH74 family could promote rapid degradation of host tissues by *C. tanaceti* during the necrotrophic phase. The expansion of the lignin-degrading AA2 family in *C. tanaceti* has the potential to assist infection of xylem vessels and thereby aid translocation of propagules to different parts of the plant and establishing secondary infections.

The conserved nature of certain pathogenicity genes, such as the secondary metabolite clusters within the destructivum complex, was evident with their presence within the syntenic blocks with *C. higginsianum*. However, only a minority of the effectors, proteases and SM backbone genes of *C. tanaceti* were among the core gene set for *Colletotrichum* spp. tested, therefore emphasizing their role in adaptation to new hosts. The species-specific effectors, singletons from the orthology analysis and the genes exclusive to *C. tanaceti* might have been horizontally transferred or be related to the host affiliation and niche specialization of *C. tanaceti*. Taken together, this inferred pathogenicity gene suite of *C. tanaceti* could be targeted in future resistance breeding and other disease management strategies for *C. tanaceti*.

Host range of *Colletotrichum tanaceti*

The proposed pathogenicity gene repertoire of *C. tanaceti* was most similar to that of pathogens with intermediate host ranges. The number of pathogenicity genes inferred from *C. tanaceti* was either similar to or less than the average for all *Colletotrichum spp*. investigated, but the overall composition was similar to *Colletotrichum* spp. which either were able to infect many species within a plant family or few species across families.

The putative pathogenicity profile of *C. tanaceti* was very distinct from that of the other destructivum complex member, *C. higginsianum*, despite the two species sharing the highest number of orthologs and having the shortest evolutionary distance. Contractions in many pathogenicity gene families in *C. tanaceti* compared to *C. higggginsianum* indicated more restricted pathogenicity in *C. tanaceti*. The most similar CAZyme pathogenicity profile to that of *C. tanaceti* was from *C. chlorophyti* which has been reported to infect herbaceous hosts such as tomato (plant family Solanaceae) and soybean (plant family Fabaceae) [82]. The similarity to *C. chlorophyti* was consistent for other gene categories such as the P450s, transporters and the overall pathogenicity profile. A homolog to the demethylase (*PDA*), which provides tolerance to the phytoalexin pisatin synthesised by *Pisum sativum* [158], was predicted in *C. tanaceti* (CTA1_6324s) which could be an indicator of the ability of *C. tanaceti* to infect Fabaceae. The composition of the SMB cluster was however, more similar to *C. orchidophilum*, another pathogen reported to infect the herbaceous, monocot plant family of Orchidaceae [159]. The similarity of the putative pathogenicity profile of *C. tanaceti* to two pathogens infecting

multiple herbaceous plant species was notable as the only known host of *C. tanaceti* is also herbaceous. Both *C. chlorophyti* and *C. orchidophilum* have been reported from multiple host species. Therefore, the putative pathogenicity gene suite of *C. tanaceti* suggests that *C. tanaceti* has the genetic ability to infect more hosts than currently recognized. If *C. tanaceti* can infect other hosts, such crops could also provide an external gene pool of inoculum for infection of pyrethrum crops increasing the evolutionary potential of the pathogen populations. Based on results of comparative analysis of pathogenicity profiles, a further hypothesis is that these alternative hosts are likely to be herbaceous plants. Future studies investigating the cross-host infectivity and pathogenicity of *C. tanaceti* are recommended.

Evolution of pathogenicity genes

Pathogenicity genes of *C. tanaceti* appear to be capable of evolving relatively rapidly. Tandem repeats such as simple sequence repeats have high mutation rates [160] and could promote frameshift mutations in adjacent genes by slipped misalignment during replication. Therefore, the significant overlap between the tandem repeats and the pathogenicity genes suggested high potential to mutate and create different pathotypes. Transposons promote insertional mutations that can either cause disruption or modification of gene expression or generate new proteins and also are major drivers of gene duplication [161]. Transposons were in close proximity to putative pathogenicity in *C. tanaceti*, such as the SMB clusters, expanded and contracted CAZymes, peptidases and effectors. The significant association of TE with pathogenicity genes were previously reported in *C. truncatum* [41] and *C. higginsianum* [15]. The RIP affected regions of the genome were also in close proximity to certain putative pathogenicity genes of *C. tanaceti*. RIP mutations can leak into nearby flanking regions causing mutations in those genes, further diversifying the pathogenicity gene repertoire of *C. tanaceti* [35]. However, unlike in *Leptosphaeria maculans* [162] RIP was not associated with the putative effectors of *C. tanaceti*. Therefore, TEs could be facilitating effector diversification in this species [41]. Although gene sparse, the A-T rich regions of the *C. tanaceti* genome contained several (n = 18) putative pathogenicity and virulence factors and many hypothetical proteins which could be functioning as effectors facilitating adaptive evolution. Small secretory proteins were identified in the A-T rich regions of *C. orbiculare* [6]. The genes in these A-T and repeat rich, gene sparse regions can evolve faster than the rest of the genome according to the "two-speed genomes" hypothesis [36].

Colletotrichum tanaceti had the highest number of rapidly evolving CAZyme families among the 17 species studied which also was indicative of the high evolutionary potential in these pathogenicity genes. Interspersed repeats were not in close proximity to the total CAZome. They were however, located significantly closer to the expanded or contracted families indicating that interspersed repeats were a major contributor to CAZyme family expansions/contractions in *C. tanaceti* by causing gene duplication (in expansions) or gene disruptions (in contractions) [135, 163]. Diversification of pathogenicity and virulence genes through repeats and RIP mutation in *C. tanaceti* result a high evolutionary potential for pathogenicity genes of this pathogen. The high evolutionary potential of pathogenicity genes may cause rapid evolution of resistance to host immune responses in existing hosts or even adaptation to new host species in *C. tanaceti*.

Genus *Colletotrichum*

Phylogenetic relationship throughout the genus was consistent with previous observations, with gloeosporioides complex members and *C. orbiculare* forming a clade separately from the destructivum, graminicola and acutatum clades [9–11]. One notable difference was in the

divergence time estimates for the divergence of *Colletotrichum* species complexes which were more ancient than reported by Liang et al [11], despite using the same calibration times. This could have been due to this study using cross-validation across 50 smoothing factors in CAFÉ as opposed to using 12 different constraints and smoothing factor combinations differences, as well as the use of the use of the different data sets.

Comparative genomic analyses emphasized the rapid evolutionary rate and the high diversity within the genus. The short time for speciation within the acutatum complex, and the fourteen *Colletotrichum* species in general, was suggestive of the high evolutionary rate within the genus with respective to the typical evolutionary rate of the fungal kingdom (0.0085 species units per Myr) [164]. The sequence similarity between *C. tanaceti* and other species of *Colletotrichum* varied widely and dropped drastically with evolutionary distance, suggesting high diversity within the genus. However, the drop in orthology was less dramatic, emphasizing the contribution of non-coding regions in generating diversity within the genus. The extent of synteny between *C. tanaceti* and *C. higginsianum* was high and very similar to the percentage synteny previously reported for the two graminicola complex species, *C. sublineola* and *C. graminicola* [165]. This suggested that even though there was high diversity within the genus, the species in the same species complex tend to share more synteny and orthology than the species between species complexes.

Evolutionary analysis of CAZyme families of different *Colletotrichum* lineages revealed an association between CAZyme families and host range. The GH131 with cellulose degrading activity was the only rapidly evolving gene family at the MRCA of *Colletotrichum* spp. suggesting a possible association of this family with speciation and host determination within the genus. Families GH43, with hemicellulose degrading activity and AA7, with gluco-oligosaccharide activity significantly expanded upon divergence of both the gloeosporioides and acutatum-complex clades, which could have broadened the host ranges of members of these two complexes, consistent with previous reports [7, 9–11].The significant expansions in pectin degrading enzyme families GH106 in gloeosporioides and GH78 in the acutatum clades could also have enabled degradation of pectin rich cell walls of young fruits [166] of these fruit-rotting species.

The most significant contractions were reported in pectin degrading families upon the divergence of the graminicola complex clade. This could have been the reason for species in this complex exclusively infecting monocot plant species considering that the pectin content of monocot cell walls is generally less than in dicots [167]. Even though this was a similar result to previous studies [6, 7, 9–11, 43], *C. orchidophilum* which is known to infect plants from monocot family Orchidaceae [168], deviated from this pattern. Gene family AA7 was rapidly evolving in many *Colletotrichum* species and could have been involved in biotransformation or detoxification of the lignocellulosic compounds [169].

In general, the overall CAZyme pathogenicity profiles of *Colletotrichum* spp. followed host range of those species rather than the taxonomy in consistence with the previous studies, [6, 7, 9–11, 43, 170]. The gloeoporioides and acutatum complex members which have broad host ranges, but are evolutionary distant, were clustered together. This could be a byproduct of the "two-speed" genome scenario in certain *Colletotrichum* spp. such as *C. orbiculare*, *C. fructicola* and *C. graminicola* [6, 11, 41] and as suggested by this study, also in *C. tanaceti*. In this scenario, the pathogenicity genes are located in repeat-rich regions, allowing them to evolve at a higher rate than the rest of the genome. This was also evident by the significant association of TE with pathogenicity genes in *C. tanaceti*, *C. truncatum*, *C. higginsianum* and *C. graminicola* [10, 15, 43]. Furthermore, in *C. fructicola*, two gene clusters that were horizontally transferred were within the rapidly evolving lineage specific regions [11]. This scenario would cause the

species with similar pathogenicity gene profiles to cluster together, despite their evolutionary distance.

Conclusion

In conclusion, a draft genome of *C. tanaceti* was used to characterize the molecular basis of pathogenicity of the species and to improve the knowledge of the evolution of the fungal genus *Colletotrichum*. *Colletotrichum tanaceti* is likely to have alternative hosts to pyrethrum. The genome of *Colletotrichum tanaceti* contains a large component of repetitive elements that may result in genome expansion and rapid generation of novel genotypes. The tendency of the pathogenicity genes to evolve rapidly was evident in genomic signals of the RIP and association of repeats and RIPs with the putative pathogenicity genes. Therefore, due to the large array of pathogenicity genes with a high evolutionary potential, *C. tanaceti* is likely to become a high-risk pathogen. Complexity of the *Colletotrichum* genus was evident with its high diversity and evolutionary rate. The significant expansions and contractions of gene families upon divergence of different lineages within the genus could be important determinants in species and species complex diversification in *Colletotrichum*. The reason for pathogenicity genes to have different clustering than the phylogeny in *Colletotrichum* could be the occurrence of "two-speed" genomes in certain species. These findings will facilitate future research in genomics and disease management of *Colletotrichum*.

Supporting information

S1 Table. GO term enrichment analysis in *C. tanaceti*.
(XLSX)

S2 Table. KEGG orthology annotations of *C. tanaceti*.
(XLSX)

S3 Table. KEGG pathway map IDs of *C. tanaceti*.
(XLSX)

S4 Table. KEGG orthology assignments of Map kinase pathway in *C. tanaceti*.
(XLSX)

S5 Table. KEGG orthology assignments of Aflatoxin biosynthesis pathway in *C. tanaceti*.
(XLSX)

S6 Table. KEGG orthology assignments of ABC transporters in *C. tanaceti*.
(XLSX)

S7 Table. Global alignment of *C. tanaceti* contigs to the *C. higginsianum* chromosomes.
(XLSX)

S8 Table. Secreted proteins of *C. tanaceti* and their conserved domains.
(XLSX)

S9 Table. Secreted effector candidates of *C. tanaceti* with homology to known effectors and conserved motifs.
(XLSX)

S10 Table. Secreted peptidases of *C. tanaceti*.
(XLSX)

S11 Table. Secreted peptidase inhibitors of *C. tanaceti*.
(XLSX)

S12 Table. Secondary metabolite biosynthetic gene cluster predictions of *C. tanaceti*.
(XLSX)

S13 Table. Conserved domains of secondary metabolite biosynthetic genes of *C. tanaceti*.
(XLSX)

S14 Table. Homologs in *C. tanaceti* to fungal cytochrome P450 database.
(XLSX)

S15 Table. Homologs in *C. tanaceti* to transporter classification database.
(XLSX)

S16 Table. Homologs to pathogen host interaction database in *C. tanaceti*.
(XLSX)

S17 Table. CAZyme family assignment of *C. tanaceti*.
(XLSX)

S18 Table. Family-wide probability values and viterbi probability values of CAZyme families across taxa.
(XLSX)

S19 Table. Expansions and contractions of CAZyme families upon divergence of different lineages.
(XLSX)

S20 Table. Statistics of CAZyme gene family evolution across taxa.
(XLSX)

S21 Table. Dinucleotide frequencies of the whole genome and the A-T rich regions of C. tanaceti.
(XLSX)

S22 Table. Genes in the AT rich region of *C. tanaceti* genome.
(XLSX)

S1 Fig. Median GC percentage and median total length (Mb) (x axis) of publicly available draft genomes representing *Colletotrichum* species (y axis).
(TIF)

S2 Fig. Circular plot showing synteny between *Colletotrichum tanaceti* contigs (numbers) mapped to the 12 individual chromosomes (NC_codes) of the *C. higginsianum* genome.
(TIF)

S3 Fig. Comparison of type of secondary metabolite biosynthetic gene clusters (gene clusters producing Terpenes, Indoles, Polyketides (PKs), Non-ribosomal peptides (NRPs_and hybrids of the above categories) in *Colletotrichum* species; hierarchical clustering performed using Euclidean distance and Ward linkage. The number of genes in each gene category is normalized using unit variance scaling. Overrepresented and underrepresented types of secondary metabolite gene clusters are represented in red to orange and blue respectively as fold standard deviations above and below the mean.
(TIF)

S4 Fig. Comparison of composition of pathogen host interaction database (PHIbase) homolog profiles (number homologs to entries in "reduced virulence", "unaffected pathogenicity", loss of pathogenicity", "effector", "lethal" and "increased virulence" categories in the PHIbase) in *Colletotrichum* and related species; hierarchical clustering performed with Euclidean distance and Ward linkage. The number of genes in each PHI category is normalized using unit variance scaling. Overrepresented and underrepresented gene categories are represented in red to orange and blue respectively as fold standard deviations above and below the mean.
(TIF)

Acknowledgments

We thank Dr Kym Pham and Dr Arthur Hsu for performing library preparation, sequencing and genome assembly which were conducted at the Melbourne Translational Genomics Platform (Department of Pathology, University of Melbourne). We also thank Botanical Resources Australia—Agricultural Services, Pty. Ltd for supporting the project. RVL received Melbourne International Fee Remission Scholarship and Melbourne International Research Scholarship from the University of Melbourne, Australia. NDY was supported by a Career Development Fellowship (CDF) from NHMRC. PKK was supported by an Early Career Fellowship (CDF) from NHMRC.

Author Contributions

Conceptualization: Ruvini V. Lelwala, Paul W. J. Taylor.

Data curation: Ruvini V. Lelwala, Pasi K. Korhonen.

Formal analysis: Ruvini V. Lelwala, Pasi K. Korhonen.

Funding acquisition: Paul W. J. Taylor.

Investigation: Ruvini V. Lelwala.

Methodology: Ruvini V. Lelwala, Pasi K. Korhonen, Neil D. Young.

Project administration: Paul W. J. Taylor.

Resources: Robin B. Gasser, Paul W. J. Taylor.

Supervision: Pasi K. Korhonen, Neil D. Young, Jason B. Scott, Peter K. Ades, Robin B. Gasser, Paul W. J. Taylor.

Validation: Pasi K. Korhonen, Neil D. Young, Jason B. Scott, Peter K. Ades, Robin B. Gasser, Paul W. J. Taylor.

Visualization: Ruvini V. Lelwala.

Writing – original draft: Ruvini V. Lelwala.

Writing – review & editing: Ruvini V. Lelwala, Pasi K. Korhonen, Neil D. Young, Jason B. Scott, Peter K. Ades, Robin B. Gasser, Paul W. J. Taylor.

References

1. Velásquez AC, Castroverde CDM, He SY. Plant-Pathogen Warfare under Changing Climate Conditions. Curr Biol. 2018; 28(10):R619–R34. https://doi.org/10.1016/j.cub.2018.03.054 PMID: 29787730
2. Doehlemann G, Ökmen B, Zhu W, Sharon A. Plant Pathogenic Fungi. Microbiology Spectrum. 2017; 5 (1). https://doi.org/10.1128/microbiolspec.FUNK-0023-2016 PMID: 28155813

3. Dean R, Van K A L, Pretorius Z A, Hammond-Kosack K E, Di P A, Spanu P D, et al. The Top 10 fungal pathogens in molecular plant pathology. Mol Plant Pathol. 2012; 13(4):414–30. https://doi.org/10.1111/j.1364-3703.2011.00783.x PMID: 22471698
4. Cannon PF, Damm U, Johnston PR, Weir BS. *Colletotrichum*–current status and future directions. Studies in Mycology. 2012; 73(1):181–213. https://doi.org/10.3114/sim0014 PMC3458418. PMID: 23136460
5. O'Connell RJ, Thon MR, Hacquard S, Amyotte SG, Kleemann J, Torres MF, et al. Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and transcriptome analyses. Nat Genet. 2012; 44(9):1060–5. http://www.nature.com/ng/journal/v44/n9/abs/ng.2372.html#supplementary-information. https://doi.org/10.1038/ng.2372 PMID: 22885923
6. Gan P, Ikeda K, Irieda H, Narusaka M, O'Connell RJ, Narusaka Y, et al. Comparative genomic and transcriptomic analyses reveal the hemibiotrophic stage shift of *Colletotrichum* fungi. New Phytol. 2013; 197(4):1236–49. https://doi.org/10.1111/nph.12085 PMID: 23252678.
7. Baroncelli R, Amby DB, Zapparata A, Sarrocco S, Vannacci G, Le Floch G, et al. Gene family expansions and contractions are associated with host range in plant pathogens of the genus *Colletotrichum*. BMC Genomics. 2016; 17(1):555. https://doi.org/10.1186/s12864-016-2917-6 PMID: 27496087
8. Hacquard S, Kracher B, Hiruma K, Münch PC, Garrido-Oter R, Thon MR, et al. Survival trade-offs in plant roots during colonization by closely related beneficial and pathogenic fungi. Nature Communications. 2016; 7:11362. https://doi.org/10.1038/ncomms11362 https://www.nature.com/articles/ncomms11362#supplementary-information. PMID: 27150427
9. Gan P, Narusaka M, Kumakura N, Tsushima A, Takano Y, Narusaka Y, et al. Genus-Wide Comparative Genome Analyses of *Colletotrichum* Species Reveal Specific Gene Family Losses and Gains during Adaptation to Specific Infection Lifestyles. *Genome Biology and Evolution*. 2016; 8(5):1467–81. https://doi.org/10.1093/gbe/evw089 PMID: 27189990
10. Rao S, Nandineni MR. Genome sequencing and comparative genomics reveal a repertoire of putative pathogenicity genes in chilli anthracnose fungus *Colletotrichum truncatum*. PLOS ONE. 2017; 12(8):e0183567. https://doi.org/10.1371/journal.pone.0183567 PMID: 28846714
11. Liang X, Wang B, Dong Q, Li L, Rollins JA, Zhang R, et al. Pathogenic adaptations of *Colletotrichum* fungi revealed by genome wide gene family evolutionary analyses. PLOS ONE. 2018; 13(4):e0196303. https://doi.org/10.1371/journal.pone.0196303 PMID: 29689067
12. Mongkolporn O, Taylor PWJ. Chili anthracnose: *Colletotrichum* taxonomy and pathogenicity. Plant Pathol. 2018; 67(6):1255–63. https://doi.org/10.1111/ppa.12850
13. Marin-Felix Y, Hernández-Restrepo M, Wingfield MJ, Akulov A, Carnegie AJ, Cheewangkoon R, et al. Genera of phytopathogenic fungi: GOPHY 2. Studies in Mycology. 2019; 92:47–133. https://doi.org/10.1016/j.simyco.2018.04.002 PMID: 29997401
14. Damm U, Sato T, Alizadeh A, Groenewald JZ, Crous PW. The *Colletotrichum dracaenophilum*, *C. magnum* and *C. orchidearum* species complexes. Studies in Mycology. 2019; 92:1–46. https://doi.org/10.1016/j.simyco.2018.04.001 PMID: 29997400
15. Dallery J-F, Lapalu N, Zampounis A, Pigné S, Luyten I, Amselem J, et al. Gapless genome assembly of *Colletotrichum higginsianum* reveals chromosome structure and association of transposable elements with secondary metabolite gene clusters. BMC Genomics. 2017; 18(1):667. https://doi.org/10.1186/s12864-017-4083-x PMID: 28851275
16. Damm U, O'Connell RJ, Groenewald JZ, Crous PW. The *Colletotrichum* destructivum species complex—hemibiotrophic pathogens of forage and field crops. Stud Mycol. 2014; 79:49–84. Epub 2014/12/11. https://doi.org/10.1016/j.simyco.2014.09.003 PMID: 25492986; PubMed Central PMCID: PMCPmc4255528.
17. Barimani M, Pethybridge SJ, Vaghefi N, Hay FS, Taylor PWJ. A new anthracnose disease of pyrethrum caused by *Colletotrichum tanaceti* sp. nov. Plant Pathol. 2013; 62(6):1248–57. https://doi.org/10.1111/ppa.12054 PMID: 91824730.
18. Duke SO, Cantrell CL, Meepagala KM, Wedge DE, Tabanca N, Schrader KK. Natural toxins for use in pest management. Toxins. 2010; 2(8):1943–62. https://doi.org/10.3390/toxins2081943 PMID: 22069667.
19. Hay FS, Gent DH, Pilkington SJ, Pearce TL, Scott JB, Pethybridge SJ. Changes in distribution and frequency of fungi associated with a foliar disease complex of pyrethrum in Australia. Plant Dis. 2015; 99(9):1227–35. https://doi.org/10.1094/PDIS-12-14-1357-RE PMID: 30695926
20. De Silva DD, Crous PW, Ades PK, Hyde KD, Taylor PWJ. Life styles of *Colletotrichum* species and implications for plant biosecurity. Fungal Biology Reviews. 2017; 31(3):155–68. https://doi.org/10.1016/j.fbr.2017.05.001.

21. Sperschneider J, Gardiner D, M., Dodds PN, Tini F, Covarelli L, Singh KB, et al. EffectorP: predicting fungal effector proteins from secretomes using machine learning. New Phytol. 2015; 210(2):743–61. https://doi.org/10.1111/nph.13794 PMID: 26680733

22. Muszewska A, Stepniewska-Dziubinska MM, Steczkiewicz K, Pawlowska J, Dziedzic A, Ginalski K. Fungal lifestyle reflected in serine protease repertoire. *Scientific Reports*. 2017; 7(1):9147. https://doi.org/10.1038/s41598-017-09644-w PMID: 28831173

23. Zhao Z, Liu H, Wang C, Xu J-R. Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. BMC Genomics. 2013; 14(1):274. https://doi.org/10.1186/1471-2164-14-274 PMID: 23617724

24. Howlett BJ. Secondary metabolite toxins and nutrition of plant pathogenic fungi. Curr Opin Plant Biol. 2006; 9(4):371–5. https://doi.org/10.1016/j.pbi.2006.05.004 PMID: 16713733

25. Zhao X, Mehrabi R, Xu J-R. Mitogen-Activated Protein Kinase Pathways and Fungal Pathogenesis. Eukaryot Cell. 2007; 6(10):1701–14. https://doi.org/10.1128/EC.00216-07 PMID: 17715363

26. Sharma KK. Fungal genome sequencing: basic biology to biotechnology. Crit Rev Biotechnol. 2016; 36(4):743–59. https://doi.org/10.3109/07388551.2015.1015959 PMID: 25721271

27. Yoder OC, Turgeon BG. Fungal genomics and pathogenicity. Curr Opin Plant Biol. 2001; 4(4):315–21. https://doi.org/10.1016/S1369-5266(00)00179-5. PMID: 11418341

28. Urban M, Cuzick A, Rutherford K, Irvine A, Pedro H, Pant R, et al. PHI-base: a new interface and further additions for the multi-species pathogen–host interactions database. Nucleic Acids Res. 2017; 45 (D1):D604–D10. https://doi.org/10.1093/nar/gkw1089 PMID: 27915230

29. Lu T, Yao B, Zhang C. DFVF: database of fungal virulence factors. Database: The Journal of Biological Databases and Curation. 2012;2012:bas032. https://doi.org/10.1093/database/bas032 PMC3478563. PMID: 23092926

30. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, et al. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. 2015; 43(W1): W237–W43. https://doi.org/10.1093/nar/gkv437 PMID: 25948579

31. Howlett BJ, Lowe RG, Marcroft SJ, van de Wouw AP. Evolution of virulence in fungal plant pathogens: exploiting fungal genomics to control plant disease. Mycologia. 2015; 107(3):441–51. Epub 2015/03/01. https://doi.org/10.3852/14-317 PMID: 25725000.

32. McDonald BA, Linde C. Pathogen population genetics, evotutionary potential, and durable resistance. Annu Rev Phytopathol. 2002; 40(1):349–79. https://doi.org/10.1146/annurev.phyto.40.120501.101443 PMID: 12147764.

33. Aylward J, Steenkamp ET, Dreyer LL, Roets F, Wingfield BD, Wingfield MJ. A plant pathology perspective of fungal genome sequencing. IMA fungus. 2017; 8(1):1–15. Epub 02/09. https://doi.org/10.5598/imafungus.2017.08.01.01 PMID: 28824836.

34. Stukenbrock EH. Evolution, selection and isolation: a genomic view of speciation in fungal plant pathogens. New Phytol. 2013; 199(4):895–907. https://doi.org/10.1111/nph.12374 PMID: 23782262

35. Raffaele S, Kamoun S. Genome evolution in filamentous plant pathogens: why bigger can be better. Nature Reviews Microbiology. 2012; 10:417. https://doi.org/10.1038/nrmicro2790 PMID: 22565130

36. Dong S, Raffaele S, Kamoun S. The two-speed genomes of filamentous pathogens: waltz with plants. Curr Opin Genet Dev. 2015; 35:57–65. https://doi.org/10.1016/j.gde.2015.09.001 PMID: 26451981

37. Testa AC, Oliver RP, Hane JK. OcculterCut: A Comprehensive Survey of AT-Rich Regions in Fungal Genomes. Genome biology and evolution. 2016; 8(6):2044–64. https://doi.org/10.1093/gbe/evw121 PMID: 27289099.

38. Cambareri EB, Jensen BC, Schabtach E, Selker EU. Repeat-induced GC to AT mutations in *Neurospora*. Science. 1989; 244(4912):1571–5. PMID: 2544994

39. Santana MF, Silva JC, Mizubuti ES, Araújo EF, Condon BJ, Turgeon BG, et al. Characterization and potential evolutionary impact of transposable elements in the genome of *Cochliobolus heterostrophus*. BMC Genomics. 2014; 15(1):536.

40. Li W-C, Huang C-H, Chen C-L, Chuang Y-C, Tung S-Y, Wang T-F. *Trichoderma reesei* complete genome sequence, repeat-induced point mutation, and partitioning of CAZyme gene clusters. Biotechnology for biofuels. 2017; 10(1):170.

41. Rao S, Sharda S, Oddi V, Nandineni MR. The Landscape of Repetitive Elements in the Refined Genome of Chilli Anthracnose Fungus *Colletotrichum truncatum*. Frontiers in Microbiology. 2018; 9 (2367). https://doi.org/10.3389/fmicb.2018.02367 PMID: 30337918

42. Crouch JA, O'Connell R, Gan P, Buiate E, Torres M, Beirn L, et al. The Genomics of *Colletotrichum*2014.

43. O'Connell RJ, Thon MR, Hacquard S, Amyotte SG, Kleemann J, Torres MF, et al. Lifestyle transitions in plant pathogenic *Colletotrichum fungi* deciphered by genome and transcriptome analyses. Nat Genet. 2012; 44. https://doi.org/10.1038/ng.2372 PMID: 22885923

44. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin. 1987; 19:11–5. citeulike-article-id:678648.

45. Kapabiosystems. Kapa Biosystems | Enzyme Solutions | Next Generation PCR 2015 [cited 2015 08/11]. Available from: https://www.kapabiosystems.com/.

46. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–20. Epub 2014/04/04. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404; PubMed Central PMCID: PMCPmc4103590.

47. Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. 2017; 33(4):574–6. https://doi.org/10.1093/bioinformatics/btw663 PMID: 27797770

48. Broad Institute. software.broadinstitute.org 2018 [cited 2016 12/12]. Available from: https://software.broadinstitute.org/software/discovar/blog/?page_id=98.

49. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015; 31(19):3210–2. https://doi.org/10.1093/bioinformatics/btv351 PMID: 26059717

50. Afgan E, Sloggett C, Goonasekera N, Makunin I, Benson D, Crowe M, et al. Genomics Virtual Laboratory: A Practical Bioinformatics Workbench for the Cloud. *PLOS ONE*. 2015; 10(10):e0140829. https://doi.org/10.1371/journal.pone.0140829 PMID: 26501966

51. Smit AFA, Robert H, Kas A, Siegel A, Gish W, Price A, et al. RepeatModeler. 1.0.5 ed. http://www.repeatmasker.org: Institute of Systems Biology; 2011.

52. Bao Z, Eddy SR. Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res. 2002; 12(8):1269–76. Epub 2002/08/15. https://doi.org/10.1101/gr.88502 PMID: 12176934; PubMed Central PMCID: PMC186642.

53. Price AL, Jones NC, Pevzner PA. *De novo* identification of repeat families in large genomes. Bioinformatics. 2005; 21 Suppl 1:i351–8. Epub 2005/06/18. https://doi.org/10.1093/bioinformatics/bti1018 PMID: 15961478.

54. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007; 35(suppl 2):W265–W8.

55. Smit AFA, Hubley R, Green P. RepeatMasker. http://www.repeatmasker.org: Institute of Systems Biology; 1996–2010.

56. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999; 27(2):573–80. Epub 1998/12/24. https://doi.org/10.1093/nar/27.2.573 PMID: 9862982; PubMed Central PMCID: PMC148217.

57. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005; 110(1–4):462–7. Epub 2005/08/12. https://doi.org/10.1159/000084979 PMID: 16093699.

58. Auyong ASM, Ford R, Taylor PWJ. Genetic transformation of *Colletotrichum truncatum* associated with anthracnose disease of chili by random insertional mutagenesis. J Basic Microbiol. 2012; 52(4):372–82. https://doi.org/10.1002/jobm.201100250 PMID: 22052577

59. Chen X, Shen G, Wang Y, Zheng X, Wang Y. Identification of *Phytophthora sojae* genes upregulated during the early stage of soybean infection. FEMS Microbiol Lett. 2007; 269(2):280–8. https://doi.org/10.1111/j.1574-6968.2007.00639.x PMID: 17263843

60. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data 2017 [cited 2017]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

61. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC bioinformatics. 2011; 12:491. Epub 2011/12/24. https://doi.org/10.1186/1471-2105-12-491 PMID: 22192575; PubMed Central PMCID: PMC3280279.

62. O'Connell RJ, Thon MR, Hacquard S, Amyotte SG, Kleemann J, Torres MF, et al. Lifestyle transitions in plant pathogenic Colletotrichum fungi deciphered by genome and transcriptome analyses. Nat Genet. 2012; 44(9):1060–5. https://doi.org/10.1038/ng.2372 PMID: 22885923

63. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013; 8(8):1494–512. Epub 2013/07/13. https://doi.org/10.1038/nprot.2013.084 PMID: 23845962; PubMed Central PMCID: PMC3875132.

64. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology. 2013; 14(4): R36. https://doi.org/10.1186/gb-2013-14-4-r36 PMID: 23618408

65. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012; 28(23):3150–2. https://doi.org/10.1093/bioinformatics/bts565 PMC3516142. PMID: 23060610

66. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22(13):1658–9. https://doi.org/10.1093/bioinformatics/btl158 PMID: 16731699

67. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006; 34(Web Server issue):W435–9. Epub 2006/07/18. https://doi.org/10.1093/nar/gkl200 PMID: 16845043; PubMed Central PMCID: PMCPmc1538822.

68. Korf I. Gene finding in novel genomes. BMC bioinformatics. 2004; 5:59. Epub 2004/05/18. https://doi.org/10.1186/1471-2105-5-59 PMID: 15144565; PubMed Central PMCID: PMC421630.

69. Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 1998; 26(4):1107–15. Epub 1998/03/21. https://doi.org/10.1093/nar/26.4.1107 PMID: 9461475; PubMed Central PMCID: PMC147337.

70. R Core Team. R: A language and environment for statistical computing.: R Foundation for Statistical Computing, Vienna, Austria.; 2015.

71. Li L, Stoeckert CJ Jr., Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003; 13(9):2178–89. https://doi.org/10.1101/gr.1224503 PMID: 12952885.

72. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10:421. Epub 2009/12/17. https://doi.org/10.1186/1471-2105-10-421 PMID: 20003500; PubMed Central PMCID: PMCPmc2803857.

73. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014; 30(9):1236–40. https://doi.org/10.1093/bioinformatics/btu031 PMC3998142. PMID: 24451626

74. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol. 2016; 428(4):726–31. https://doi.org/10.1016/j.jmb.2015.11.006 PMID: 26585406

75. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017; 45(Database issue):D353–D61. https://doi.org/10.1093/nar/gkw1092 PMC5210567. PMID: 27899662

76. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*. 2008; 4:44. https://doi.org/10.1038/nprot.2008.211 https://www.nature.com/articles/nprot.2008.211#supplementary-information. PMID: 19131956

77. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37(1):1–13. https://doi.org/10.1093/nar/gkn923 PMC2615629. PMID: 19033363

78. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. *Genome Biology*. 2004; 5(2):R12–R. https://doi.org/10.1186/gb-2004-5-2-r12 PMC395750. PMID: 14759262

79. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLoS Comp Biol. 2018; 14(1):e1005944–e. https://doi.org/10.1371/journal.pcbi.1005944 PMID: 29373581.

80. Soderlund C, Nelson W, Shoemaker A, Paterson A. SyMAP: A system for discovering and viewing syntenic regions of FPC maps. Genome Res. 2006; 16(9):1159–68. https://doi.org/10.1101/gr.5396706 PMC1557773. PMID: 16951135

81. Soderlund C, Bomhoff M, Nelson WM. SyMAP v3.4: a turnkey synteny system with application to plant genomes. Nucleic Acids Res. 2011; 39(10):e68. Epub 2011/03/15. https://doi.org/10.1093/nar/gkr123 PMID: 21398631; PubMed Central PMCID: PMCPMC3105427.

82. Gan P, Narusaka M, Tsushima A, Narusaka Y, Takano Y, Shirasu K. Draft Genome Assembly of *Colletotrichum chlorophyti*, a Pathogen of Herbaceous Plants. Genome Announc. 2017; 5(10). Epub 2017/03/11. https://doi.org/10.1128/genomeA.01733-16 PMID: 28280027; PubMed Central PMCID: PMCPMC5347247.

83. Baroncelli R, Sreenivasaprasad S, Sukno SA, Thon MR, Holub E. Draft Genome Sequence of *Colletotrichum acutatum* Sensu Lato (*Colletotrichum fioriniae*). Genome announcements. 2014; 2(2): e00112–14. https://doi.org/10.1128/genomeA.00112-14 PMID: 24723700.

84. Alkan N, Meng X, Friedlander G, Reuveni E, Sukno S, Sherman A, et al. Global Aspects of pacC Regulation of Pathogenicity Genes in *Colletotrichum gloeosporioides* as Revealed by Transcriptome Analysis. Mol Plant-Microbe Interact. 2013; 26(11):1345–58. https://doi.org/10.1094/MPMI-03-13-0080-R PMID: 23902260

85. Baroncelli R, Sukno SA, Sarrocco S, Cafa G, Le Floch G, Thon MR. Whole-Genome Sequence of the Orchid Anthracnose Pathogen *Colletotrichum orchidophilum*. Mol Plant Microbe Interact. 2018; 31 (10):979–81. Epub 2018/04/14. https://doi.org/10.1094/MPMI-03-18-0055-A PMID: 29649963.

86. Baroncelli R, Sanz-Martín JM, Rech GE, Sukno SA, Thon MR. Draft Genome Sequence of *Colletotrichum sublineola*, a Destructive Pathogen of Cultivated Sorghum. Genome Announcements. 2014; 2 (3):e00540–14. https://doi.org/10.1128/genomeA.00540-14 PMC4056296. PMID: 24926053

87. Klosterman SJ, Subbarao KV, Kang S, Veronese P, Gold SE, Thomma BPHJ, et al., editors. Comparative genomics of the plant vascular wilt pathogens, *Verticillium dahliae* and *Verticillium albo-atrum*2010.

88. Staats M, van Kan JAL. Genome update of *Botrytis cinerea* strains B05.10 and T4. Eukaryot Cell. 2012; 11(11):1413–4. https://doi.org/10.1128/EC.00164-12 PMID: 23104368.

89. Nowrousian M, Stajich JE, Chu M, Engh I, Espagne E, Halliday K, et al. De novo assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. PLoS Genet. 2010; 6(4):e1000891. Epub 2010/04/14. https://doi.org/10.1371/journal.pgen.1000891 PMID: 20386741; PubMed Central PMCID: PMCPMC2851567.

90. Ma L-J, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-J, Di Pietro A, et al. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature. 2010; 464:367. https://doi.org/10.1038/nature08850 https://www.nature.com/articles/nature08850#supplementary-information. PMID: 20237561

91. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002; 30(7):1575–84. PMC101833. https://doi.org/10.1093/nar/30.7.1575 PMID: 11917018

92. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013; 30(4):772–80. https://doi.org/10.1093/molbev/mst010 PMC3603318. PMID: 23329690

93. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25(15):1972–3. https://doi.org/10.1093/bioinformatics/btp348 PMC2712344. PMID: 19505945

94. Kück P, Longo GC. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in Zoology*. 2014; 11(1):81. https://doi.org/10.1186/s12983-014-0081-x PMID: 25426157

95. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011; 27(8):1164–5. https://doi.org/10.1093/bioinformatics/btr088 PMID: 21335321

96. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30(9):1312–3. https://doi.org/10.1093/bioinformatics/btu033 PMC3998144. PMID: 24451623

97. Paradis E, Claude J, Strimmer K. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics. 2004; 20:289. https://doi.org/10.1093/bioinformatics/btg412 PMID: 14734327

98. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2018.

99. Sanderson MJ. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics. 2003; 19(2):301–2. https://doi.org/10.1093/bioinformatics/19.2.301 PMID: 12538260

100. Sanderson MJ. Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach. Mol Biol Evol. 2002; 19(1):101–9. https://doi.org/10.1093/oxfordjournals.molbev.a003974 PMID: 11752195

101. Nash SG. A survey of truncated-Newton methods. *Journal of Computational and Applied Mathematics*. 2000; 124(1):45–59. https://doi.org/10.1016/S0377-0427(00)00426-X.

102. Beimforde C, Feldberg K, Nylinder S, Rikkinen J, Tuovila H, Dörfelt H, et al. Estimating the Phanerozoic history of the Ascomycota lineages: Combining fossil and molecular data. Mol Phylogen Evol. 2014; 78:386–98. https://doi.org/10.1016/j.ympev.2014.04.024.

103. Meinken J, Asch DK, Neizer-Ashun KA, Chang GH, Cooper CRJ, Min XJ. FunSecKB2: A fungal protein subcellular location knowledgebase2014. 1–17 p.

104. Nielsen H. Predicting Secretory Proteins with SignalP. In: Kihara D, editor. Protein Function Prediction: Methods and Protocols. New York, NY: Springer New York; 2017. p. 59–73.

105. Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res. 2007; 35(Web Server issue):W429–W32. https://doi.org/10.1093/nar/gkm256 PMC1933244. PMID: 17483518

106. Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 2007; 35(Web Server issue):W585–W7. https://doi.org/10.1093/nar/gkm259 PMC1933216. PMID: 17517783

107. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. J Mol Biol. 2001; 305(3):567–80. https://doi.org/10.1006/jmbi.2000.4315 PMID: 11152613

108. de Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 2006; 34(Web Server issue):W362–W5. https://doi.org/10.1093/nar/gkl124 PMC1538847. PMID: 16845026

109. Pierleoni A, Martelli PL, Casadio R. PredGPI: a GPI-anchor predictor. BMC Bioinformatics. 2008; 9 (1):392. https://doi.org/10.1186/1471-2105-9-392 PMID: 18811934

110. Nguyen Ba AN, Pogoutse A, Provart N, Moses AM. NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. BMC Bioinformatics. 2009; 10:202–. https://doi.org/10.1186/1471-2105-10-202 PMC2711084. PMID: 19563654

111. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Res. 2018; 46(D1):D624–D32. https://doi.org/10.1093/nar/gkx1134 PMID: 29145643

112. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol. 2010; 47(9):736–41. https://doi.org/10.1016/j.fgb.2010.06.003 PMID: 20554054

113. Park J, Lee S, Choi J, Ahn K, Park B, Park J, et al. Fungal cytochrome P450 database. BMC Genomics. 2008; 9:402–. https://doi.org/10.1186/1471-2164-9-402 PMC2542383. PMID: 18755027

114. Saier MH, Reddy VS, Tsu BV, Ahmed MS, Li C, Moreno-Hagelsieb G. The Transporter Classification Database (TCDB): recent advances. Nucleic Acids Res. 2016; 44(Database issue):D372–D9. https://doi.org/10.1093/nar/gkv1103 PMC4702804. PMID: 26546518

115. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2012; 40(Web Server issue):W445–W51. https://doi.org/10.1093/nar/gks479 PMC3394287. PMID: 22645317

116. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics. 2010; 11(1):431. https://doi.org/10.1186/1471-2105-11-431 PMID: 20718988

117. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 2006; 22(10):1269–71. https://doi.org/10.1093/bioinformatics/btl097 PMID: 16543274

118. Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3. Mol Biol Evol. 2013; 30 (8):1987–97. https://doi.org/10.1093/molbev/mst100 PMID: 23709260

119. Metsalu T, Vilo J. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Res. 2015; 43(Web Server issue):W566–W70. https://doi.org/10.1093/nar/gkv468 PMC4489295. PMID: 25969447

120. Hane JK, Oliver RP. RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. BMC Bioinformatics. 2008; 9(1):478. https://doi.org/10.1186/1471-2105-9-478 PMID: 19014496

121. Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics. 2016; 32(2):289–91. https://doi.org/10.1093/bioinformatics/btv562 PMID: 26424858

122. Derbyshire MK, Lanczycki CJ, Bryant SH, Marchler-Bauer A. Annotation of functional sites with the Conserved Domain Database. Database. 2012;2012:bar058–bar. https://doi.org/10.1093/database/bar058 PMID: 22434827

123. de Jonge R, Peter van Esse H, Kombrink A, Shinya T, Desaki Y, Bours R, et al. Conserved Fungal LysM Effector Ecp6 Prevents Chitin-Triggered Immunity in Plants. *Science*. 2010; 329(5994):953–5. https://doi.org/10.1126/science.1190859 PMID: 20724636

124. Saitoh H, Fujisawa S, Mitsuoka C, Ito A, Hirabuchi A, Ikeda K, et al. Large-Scale Gene Disruption in *Magnaporthe oryzae* Identifies MC69, a Secreted Protein Required for Infection by Monocot and Dicot Fungal Pathogens. PLoS Path. 2012; 8(5):e1002711. https://doi.org/10.1371/journal.ppat.1002711 PMID: 22589729

125. Vargas WA, Sanz-Martín JM, Rech GE, Armijos-Jaramillo VD, Rivera LP, Echeverria MM, et al. A Fungal Effector With Host Nuclear Localization and DNA-Binding Properties Is Required for Maize Anthracnose Development. Mol Plant-Microbe Interact. 2015; 29(2):83–95. https://doi.org/10.1094/MPMI-09-15-0209-R PMID: 26554735

126. Kleemann J, Rincon-Rivera LJ, Takahara H, Neumann U, van Themaat EVL, van der Does HC, et al. Sequential Delivery of Host-Induced Virulence Effectors by Appressoria and Intracellular Hyphae of the Phytopathogen *Colletotrichum higginsianum*. PLoS Path. 2012; 8(4):e1002643. https://doi.org/10.1371/journal.ppat.1002643 PMID: 22496661

127. Aboukhaddour R, Kim YM, Strelkov SE. RNA-mediated gene silencing of ToxB in Pyrenophora tritici-repentis. Mol Plant Pathol. 2012; 13(3):318–26. https://doi.org/10.1111/j.1364-3703.2011.00748.x PMID: 21980935

128. Mosquera G, Giraldo MC, Khang CH, Coughlan S, Valent B. Interaction Transcriptome Analysis Identifies Magnaporthe oryzae BAS1-4 as Biotrophy-Associated Secreted Proteins in Rice Blast Disease. *The Plant Cell*. 2009; 21(4):1273–90. https://doi.org/10.1105/tpc.107.055228 PMC2685627. PMID: 19357089

129. Sonah H, Deshmukh RK, Bélanger RR. Computational Prediction of Effector Proteins in Fungi: Opportunities and Challenges. *Frontiers in Plant Science*. 2016; 7:126. https://doi.org/10.3389/fpls.2016.00126 PMC4751359. PMID: 26904083

130. Freitag M, Williams RL, Kothe GO, Selker EU. A cytosine methyltransferase homologue is essential for repeat-induced point mutation in Neurospora crassa. Proceedings of the National Academy of Sciences. 2002; 99(13):8802–7. https://doi.org/10.1073/pnas.132212899 PMID: 12072568

131. Kouzminova E, Selker EU. dim-2 encodes a DNA methyltransferase responsible for all known cytosine methylation in Neurospora. The EMBO journal. 2001; 20(15):4309–23. https://doi.org/10.1093/emboj/20.15.4309 PMID: 11483533.

132. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011; 12(1):491. https://doi.org/10.1186/1471-2105-12-491 PMID: 22192575

133. Galagan JE, Henn MR, Ma LJ, Cuomo CA, Birren B. Genomics of the fungal kingdom: insights into eukaryotic biology. Genome Res. 2005; 15(12):1620–31. Epub 2005/12/13. https://doi.org/10.1101/gr.3767105 PMID: 16339359.

134. Karaoglu H, Lee CMY, Meyer W. Survey of Simple Sequence Repeats in Completed Fungal Genomes. Mol Biol Evol. 2005; 22(3):639–49. https://doi.org/10.1093/molbev/msi057 PMID: 15563717

135. Castanera R, López-Varas L, Borgognone A, LaButti K, Lapidus A, Schmutz J, et al. Transposable elements versus the fungal Genome: impact on whole-genome architecture and transcriptional profiles. PLoS Genet. 2016; 12(6):e1006108. https://doi.org/10.1371/journal.pgen.1006108 PMID: 27294409

136. Pellicer J, Hidalgo O, Dodsworth S, Leitch JI. Genome Size Diversity and Its Impact on the Evolution of Land Plants. Genes. 2018; 9(2). https://doi.org/10.3390/genes9020088 PMID: 29443885

137. Nieuwenhuis BPS, James TY. The frequency of sex in fungi. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2016; 371(1706):20150540. https://doi.org/10.1098/rstb.2015.0540 PMID: 27619703.

138. Crouch JA, Glasheen BM, Giunta MA, Clarke BB, Hillman BI. The evolution of transposon repeat-induced point mutation in the genome of Colletotrichum cereale: reconciling sex, recombination and homoplasy in an ''asexual" pathogen. Fungal Genet Biol. 2008; 45(3):190–206. Epub 2007/10/27. https://doi.org/10.1016/j.fgb.2007.08.004 PMID: 17962053.

139. Kulkarni RD, Kelkar HS, Dean RA. An eight-cysteine-containing CFEM domain unique to a group of fungal membrane proteins. Trends Biochem Sci. 2003; 28(3):118–21. https://doi.org/10.1016/S0968-0004(03)00025-2 PMID: 12633989

140. Zhu W, Wei W, Wu Y, Zhou Y, Peng F, Zhang S, et al. BcCFEM1, a CFEM Domain-Containing Protein with Putative GPI-Anchored Site, Is Involved in Pathogenicity, Conidial Production, and Stress Tolerance in Botrytis cinerea. *Frontiers in Microbiology*. 2017; 8(1807). https://doi.org/10.3389/fmicb.2017.01807 PMID: 28979251

141. Kim K-T, Jeon J, Choi J, Cheong K, Song H, Choi G, et al. Kingdom-Wide Analysis of Fungal Small Secreted Proteins (SSPs) Reveals their Potential Role in Host Association. *Frontiers in Plant Science*. 2016; 7:186. https://doi.org/10.3389/fpls.2016.00186 PMC4759460. PMID: 26925088

142. Selin C, de Kievit TR, Belmonte MF, Fernando WGD. Elucidating the Role of Effectors in Plant-Fungal Interactions: Progress and Challenges. Frontiers in Microbiology. 2016; 7:600. https://doi.org/10.3389/fmicb.2016.00600 PMC4846801. PMID: 27199930

143. Viaud MC, Balhadère PV, Talbot NJ. A *Magnaporthe grisea* Cyclophilin Acts as a Virulence Determinant during Plant Infection. *The Plant Cell*. 2002; 14(4):917–30. https://doi.org/10.1105/tpc.010389 PMC150692. PMID: 11971145

144. Baccelli I. Cerato-platanin family proteins: one function for multiple biological roles? *Frontiers in Plant Science*. 2014; 5:769. https://doi.org/10.3389/fpls.2014.00769 PMC4284994. PMID: 25610450

145. Bayry J, Aimanianda V, Guijarro JI, Sunde M, Latgé J-P. Hydrophobins—Unique Fungal Proteins. PLoS Path. 2012; 8(5):e1002700. https://doi.org/10.1371/journal.ppat.1002700 PMC3364958. PMID: 22693445

146. Rao MB, Tanksale AM, Ghatge MS, Deshpande VV. Molecular and biotechnological aspects of microbial proteases. Microbiology and molecular biology reviews: MMBR. 1998; 62(3):597–635. PMID: 9729602.

147. Jashni MK, Mehrabi R, Collemare J, Mesarich CH, de Wit PJGM. The battle in the apoplast: further insights into the roles of proteases and their inhibitors in plant–pathogen interactions. *Frontiers in Plant Science*. 2015; 6:584. https://doi.org/10.3389/fpls.2015.00584 PMC4522555. PMID: 26284100

148. Kroken S, Glass NL, Taylor JW, Yoder OC, Turgeon BG. Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. *Proceedings of the National Academy of Sciences*. 2003; 100(26):15670–5. https://doi.org/10.1073/pnas.2532165100 PMID: 14676319

149. Kubo Y, Furusawa I. Melanin Biosynthesis. In: Cole GT, Hoch HC, editors. The fungal spore and disease initiation in plants and animals. Boston, MA: Springer US; 1991. p. 205–18.

150. Kubo Y, Takano Y, Endo N, Yasuda N, Tajima S, Furusawa I. Cloning and structural analysis of the melanin biosynthesis gene *SCD1* encoding scytalone dehydratase in *Colletotrichum lagenarium*1997. 4340–4 p.

151. Yamada N, Motoyama T, Nakasako M, Kagabu S, Kudo T, Yamaguchi I. Enzymatic Characterization of Scytalone Dehydratase Val75Met Variant Found in Melanin Biosynthesis Dehydratase Inhibitor (MBI-D) Resistant Strains of the Rice Blast Fungus. Biosci, Biotechnol, Biochem. 2004; 68(3):615–21. https://doi.org/10.1271/bbb.68.615 PMID: 15056895

152. Becher R, Wirsel SGR. Fungal cytochrome P450 sterol 14α-demethylase (CYP51) and azole resistance in plant and human pathogens. Appl Microbiol Biotechnol. 2012; 95(4):825–40. https://doi.org/10.1007/s00253-012-4195-9 PMID: 22684327

153. Takano Y, Kikuchi T, Kubo Y, Hamer JE, Mise K, Furusawa I. The *Colletotrichum lagenarium* MAP Kinase Gene *CMK1* regulates diverse aspects of fungal pathogenesis. Mol Plant-Microbe Interact. 2000; 13(4):374–83. https://doi.org/10.1094/MPMI.2000.13.4.374 PMID: 10755300

154. Tsuji G, Fujii S, Tsuge S, Shiraishi T, Kubo Y. The *Colletotrichum lagenarium* Ste12-Like Gene *CST1* is essential for appressorium penetration. Mol Plant-Microbe Interact. 2003; 16(4):315–25. https://doi.org/10.1094/MPMI.2003.16.4.315 PMID: 12744460

155. Kojima K, Kikuchi T, Takano Y, Oshiro E, Okuno T. The mitogen-activated protein kinase gene *MAF1* is essential for the early differentiation phase of appressorium formation in *Colletotrichum lagenarium*. Mol Plant-Microbe Interact. 2002; 15(12):1268–76. https://doi.org/10.1094/MPMI.2002.15.12.1268 PMID: 12481999

156. Sakaguchi A, Tsuji G, Kubo Y. A Yeast STE11 Homologue *CoMEKK1* is essential for pathogenesis-related morphogenesis in *Colletotrichum orbiculare*. Mol Plant-Microbe Interact. 2010; 23(12):1563–72. https://doi.org/10.1094/MPMI-03-10-0051 PMID: 21039273

157. Liu P, Stajich JE. Characterization of the carbohydrate Binding Module 18 gene family in the amphibian pathogen *Batrachochytrium dendrobatidis*. Fungal Genet Biol. 2015; 77:31–9. https://doi.org/10.1016/j.fgb.2015.03.003 PMID: 25819009

158. Delserone LM, McCluskey K, Matthews DE, Vanetten HD. Pisatin demethylation by fungal pathogens and nonpathogens of pea: association with pisatin tolerance and virulence. Physiol Mol Plant Pathol. 1999; 55(6):317–26. https://doi.org/10.1006/pmpp.1999.0237.

159. Baroncelli R, Sukno S, Sarrocco S, Cafà G, Le Floch G, Thon MR. Whole-genome sequence of the orchid anthracnose pathogen Colletotrichum orchidophilum. Mol Plant-Microbe Interact. 2018. https://doi.org/10.1094/MPMI-03-18-0055-A PMID: 29649963

160. Xu X, Peng M, Fang Z, Xu X. The direction of microsatellite mutations is dependent upon allele length. Nat Genet. 2000; 24:396. https://doi.org/10.1038/74238 PMID: 10742105

161. Chia N, Goldenfeld N. Dynamics of gene duplication and transposons in microbial genomes following a sudden environmental change. Physical Review E. 2011; 83(2):021906. https://doi.org/10.1103/PhysRevE.83.021906 PMID: 21405862

162. Rouxel T, Grandaubert J, Hane JK, Hoede C, Van de Wouw AP, Couloux A, et al. Effector diversification within compartments of the Leptosphaeria maculans genome affected by Repeat-Induced Point mutations. Nature communications. 2011; 2:202. https://doi.org/10.1038/ncomms1189 PMID: 21326234

163. Seidl MF, Thomma BPHJ. Transposable elements direct the coevolution between plants and microbes. Trends Genet. 2017; 33(11):842–51. https://doi.org/10.1016/j.tig.2017.07.003 PMID: 28800915

164. Scholl JP, Wiens JJ. Diversification rates and species richness across the Tree of Life. *Proceedings of the Royal Society B*: *Biological Sciences*. 2016; 283(1838). https://doi.org/10.1098/rspb.2016.1334 PMID: 27605507

165. Buiate EAS, Xavier KV, Moore N, Torres MF, Farman ML, Schardl CL, et al. A comparative genomic analysis of putative pathogenicity genes in the host-specific sibling species *Colletotrichum graminicola* and *Colletotrichum sublineola*. BMC Genomics. 2017; 18:1–24. https://doi.org/10.1186/s12864-016-3406-7 PMID: 120657703.

166. Brummell DA. Cell wall disassembly in ripening fruit. Funct Plant Biol. 2006; 33(2):103–19. https://doi.org/10.1071/FP05234.

167. Pattathil S, Hahn MG, Dale BE, Chundawat SPS. Insights into plant cell wall structure, architecture, and integrity using glycome profiling of native and AFEXTM-pre-treated biomass. J Exp Bot. 2015; 66 (14):4279–94. https://doi.org/10.1093/jxb/erv107 PMID: 25911738

168. Damm U, Cannon PF, Woudenberg JHC, Crous PW. The *Colletotrichum* acutatum species complex. *Studies in Mycology*. 2012; 73(1):37–113. https://doi.org/10.3114/sim0010 PMC3458416. PMID: 23136458

169. Levasseur A, Drula E, Lombard V, Coutinho PM, Henrissat B. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnology for Biofuels*. 2013; 6(1):41. https://doi.org/10.1186/1754-6834-6-41 PMID: 23514094

170. Kroken S, Glass NL, Taylor JW, Yoder OC, Turgeon BG. Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. Proceedings of the National Academy of Sciences. 2003; 100(26):15670. https://doi.org/10.1073/pnas.2532165100 PMID: 14676319